Energy Efficient Cloud Computing Based Radio Access Networks in 5G

Design and Evaluation of an Energy Aware 5G Cloud Radio Access Networks Framework Using Base Station Sleeping, Cloud Computing Based Workload Consolidation and Mobile Edge Computing

Tshiamo SIGWELE

Submitted for the Degree of

Doctor of Philosophy

Faculty of Engineering and Informatics

University of Bradford

2017

ABSTRACT

Name: Tshiamo Sigwele

Thesis Title: Energy Efficient Cloud Computing Based Radio Access Networks in 5G.

Thesis Sub-Title: Design and evaluation of an energy aware 5G cloud radio access networks framework using base station sleeping, cloud computing based workload consolidation and mobile edge computing.

Keywords: 5G, Base Station Sleeping, Cloud Computing, Cloud Radio Access Networks, Energy Efficiency, Heterogeneous Networks, Mobile Edge Computing, Virtual Machine Placement, Virtualisation.

Fifth Generation (5G) cellular networks will experience a thousand-fold increase in data traffic with over 100 billion connected devices by 2020. In order to support this skyrocketing traffic demand, smaller base stations (BSs) are deployed to increase capacity. However, more BSs increase energy consumption which contributes to operational expenditure (OPEX) and CO_2 emissions. Also, an introduction of a plethora of 5G applications running in the mobile devices cause a significant amount of energy consumption in the mobile devices.

This thesis presents a novel framework for energy efficiency in 5G cloud radio access networks (C-RAN) by leveraging cloud computing technology. Energy efficiency is achieved in three ways; (i) at the radio side of H-C-RAN (Heterogeneous C-RAN), a dynamic BS switching off algorithm is proposed to minimise energy consumption while maintaining Quality of Service (QoS), (ii) in the BS cloud, baseband workload consolidation schemes are proposed based on simulated annealing and genetic algorithms to minimise energy consumption in the cloud, where also advanced fuzzy based admission control with pre-emption is implemented to improve QoS and resource utilisation (iii)

at the mobile device side, Mobile Edge Computing (MEC) is used where computer intensive tasks from the mobile device are executed in the MEC server in the cloud. The simulation results show that the proposed framework effectively reduced energy consumption by up to 48% within RAN and 57% in the mobile devices, and improved network energy efficiency by a factor of 10, network throughput by a factor of 2.7 and resource utilisation by 54% while maintaining QoS.

ACKNOWLEDGEMENTS

The writing of a thesis for the degree of Doctor of Philosophy is a lonely and isolating experience, yet not possible without personal and practical support from various people. I would like to express my thanks to the people who have been very helpful to me during this time.

First and foremost, I would like to thank our Lord *Jesus Christ* for giving me the ability to undertake this task. I also wish to express my sincere gratitude to my PhD supervisors, Dr Prashant Pillai and Prof Yim Fun Hu for their encouragement, advice and support during the duration of this PhD research work, who complemented each other wonderfully well.

In addition, I would like to express my deepest gratitude to my dear mother for her constant love, encouragement and cooperation in the duration of my study.

During my PhD, I went through a very tough medical sickness, which made me to call upon *Jesus Christ* who healed me through his anointed servants Prophet Alec Thomas in Botswana through his anointed prayers, and through his servant Bishop Dag Heward Mills through his anointed books *How to be born again and avoid Hell, Demons and how to deal with them* and *The art of leadership* and many more. I thank Bishop Dag Heward Mills, the founder of the church Mustard Seed Chapel International (MSCI) in which I gave my life to *Jesus Christ.* I also thank Reverend Steve Poku, Pastor Sunday Raji and Pastor Emmanuel Blay from MSCI branches who have completely fulfilled *Jeremiah 3:15* for me from the bible.

Hebrews 9:27 says "And as it is appointed unto men once to die, but after this the judgment" so remember your Creator while you still have life.

DEDICATION

To my Father in the Lord, Jesus Christ of Nazareth

And

To my Parents.

TABLE OF CONTENTS

CONTENTS **PAGE**

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

1G	**First Generation Cellular System**
2G	Second Generation Cellular System
3G	Third Generation Cellular System
3GPP	3rd Generation Partnership Project
4G	Fourth Generation Cellular System
5G	Fifth Generation Cellular System
5GIA	5G Infrastructure Associate
5GMF	5G Mobile Communication Promotion Forum
5GPPP	5G Infrastructure Public Private Partnership
5GrEEn	Towards Green 5G Mobile Networks
AMPS	Advanced Mobile Phone System
API	Application Programming Interface
ARPU	Average Revenue Per User
ASIC	Application Specific Integrated Circuits
AWS	Amazon Web Service
BBU	Baseband Unit
BS	Base Station
CAC	Call Admission Control
CAGR	Compound Annual Growth Rate
CAPEX	Capital Expenditure
CDMA	Code Division Multiple Access
CO_2	Carbon Dioxide
CoMP	Cooperative Multi-Point
CPRI	Common Public Radio Interface
CPU	Central Processing Unit
CQI	Channel Quality Indicator
C-RAN	Cloud Radio Access Networks
CRE	Cell Range Expansion
CSI	Channel State Information
D2D	Device to Device Communication
DPD	Digital Pre-Distortion

DSP	Digital Signal Processors
EARTH	Energy Aware Radio and Networks Technologies
EC2	Amazon Elastic Compute Cloud
EDGE	Enhanced Data Rates For GSM Evolution
EECO	Energy Efficient Computation Offloading
EEG	Electroencephalogram
eMBB	Enhanced Mobile Broadband
EMCC	Energy Efficient MEC Scheme In 5G H-C-RAN
eNodeB	Evolved NodeB
FBP	Full Bin Packing
FCC	Federal Communications Commission
FF	First Fit
FFD	First Fit Decreasing
FFT	Fast Fourier Transform
F-PDSO	Fixed Progressive Dynamic Switching Off Algorithm
FPGA	Field Programmable Gate Arrays
GA	Genetic Algorithm
GBR	Guaranteed Bit Rate
GCC	Global Cloud Controller
GCE	Google Compute Engine
GOPS	Giga Operations Per Seconds
GPP	General Purpose Processors
GPRS	General Packet Radio Service
GSM	Global System For Mobile Communications
H-C-RAN	Hetnet Cloud Radio Access Networks
HetNets	Heterogeneous Networks
HW	Hardware
ICI	Inter Cell Interference
ICT	Information and Communications Technology

IEEE	Institute of Electrical and Electronic Engineering
IETF	Internet Engineering Task Force
IMT	International Mobile Telecommunications
IoT	Internet of Things
IP	Internet Protocol
ITU	International Telecommunication Union
LoS	Line of Sight
LTE	Long Term Evolution
LTE-A	Long Term Evolution Advanced
M2M	Machine-To-Machine
MAC	Medium Access Control
MCC	Mobile Cloud Computing
MCS	Modulation And Coding Schemes
MEC	Mobile Edge Computing
MIMO	Multiple-Input Multiple-Output
MIPS	Million Instructions Per Second
MME	Mobility Management Entity
mMTC	Massive Machine Type Communication
mmW	Millimetre Waves
MOPTS	Million Operation per Time Slot
MRRH	Macro Remote Radio Head
MVCE	Mobile Virtual Centre of Excellence's Green Radio
NASDAQ	National Association of Securities Dealers Automated Quotations
NF	Next Fit
NFV	Network Function Virtualisation
NGMNA	Next Generation Mobile Networks Alliance
NIST	National Institute of Standards and Technology
Non-GBR	Not Guaranteed Bit Rate
NRT	Non Real Time

NTT	Nippon Telephone And Telegraph
O&M	Operation and Management
OBSAI	Open Base Station Architecture Initiative
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OPEX	Operational Expenditure
OS	Operating System
PA	Power Amplifier
PDA	Personal Digital Assistant
PDCP	Packet Data Convergence Protocol
PM	Physical Machine
PRB	Physical Resource Block
PRRH	Pico Remote Radio Head
QAM	Quadrature Amplitude Modulation
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Networks
RAT	Radio Access Technology
RF	Radio Frequency
RLC	Radio Link Control
RRC	Radio Resource Control
RRH	Remote Radio Head
RRM	Radio Resource Management
RSRP	Reference Signal Received Power
RT	Real Time Service
SA	Simulated Annealing
SC-FDMA	Single-Carrier Frequency Division Multiple Access
SDN	Software Defined Networks
SINR	Signal to Interference plus Noise Ratio
SLA	Service Level Agreements
SMS	Short Messaging Service

SoC	State of Charge
SON	Self-Organizing Networks
SW	Software
TACS	Total Access Communication System
TCO	Total Cost of Ownership
TCP	Transmission Control Protocol
TDMA	Time Division Multiple Access
UE	User Equipment
UL/DL	Uplink/Downlink
UMTS	Universal Mobile Telecommunications System
uRLLC	Ultra-Reliable And Low Latency Communications
vBBU	Virtual Baseband Unit
VBR	Variable Bit Rate
VBS	Virtual Base Station
VLM	Virtual Baseband Unit Live Migration
VM	Virtual Machine
VMM	Virtual Machine Monitor
VMP	Virtual Machine Placement
VP	vBBU to GPP mapping
WiFi	Wireless Fidelity
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Networks

1 Introduction

1.1 Overview

Recently, the number of connected devices have grown into billions and today mobile operators are facing the serious challenge of ever increasing demand of high data rates. For example, Huawei Technologies envisages that 100 billion devices will be connected to the internet by the year 2020 [1]. This will cause a surge in global mobile voice and data traffic. This tremendous traffic growth is due to the introduction of smart phones (Galaxy S7, IPhone, etc.), tablets (IPad, Galaxy tab, etc.), digital book readers (Kindle, etc.) and gaming consoles spawning a raft of data intensive applications, Internet of Things (IoT) and machine-to-machine (M2M) connections. As a result, next-generation mobile communication networks such as Fifth Generation (5G) have received exceptional expectations with targets to increase 1000 fold capacity, 100 times data rate, and millisecond-level delay [2] as such, there is the need to meet capacity demands. There have been three different approaches for future wireless systems to achieve much higher throughput as shown in Figure 1-1 [3].

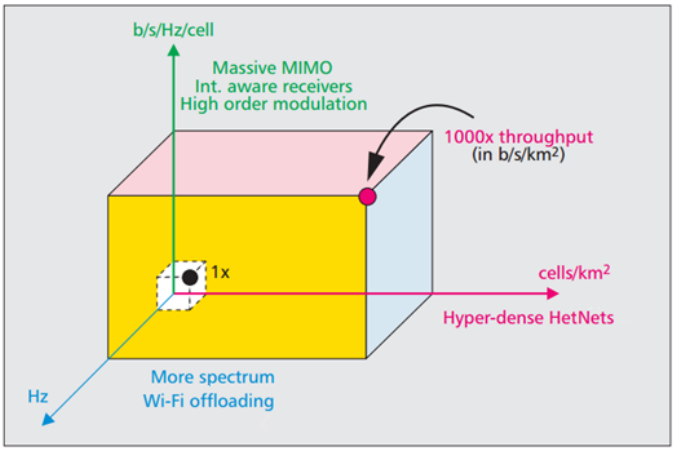

Figure 1-1. The 1000 times capacity challenge in three domains [3].

The first approach is higher utilisation of spectrum (bits per second per Hertz per cell) including massive multiple-input multiple-output (MIMO) and high modulation orders in given frequency resources per cell which is quite saturated as recent results show that at least point-to-point link throughput is very close to the theoretical limits. The second approach involves the utilisation of more bandwidth (Hertz) like Wi-Fi-offloading which is a very costly solution unless devices can utilise additional radio access technologies (RAT) for unlicensed bands with seamless aggregation and offloading. The final and probably one of the most promising frontiers to achieve the goal is to increase cells per square kilometre by deploying hyper dense cells of different types in a given area called Heterogeneous Networks (HetNets), whose goal is to maximise the utilisation of existing spectrum. In HetNets, macro cells ensure the basic coverage to meet the demand of low speed services and are overlaid with several low power nodes small cells (femto, pico, micro, relay nodes) to extend macro cell coverage and guaranteeing hotspot (airport, mall, stadium) coverage for capacity enhancement. HetNets are motivated by rethinking the network deployment principle by bringing the network closer to the user, that is, the user transmission is in proximity to the BS. However, deploying more base stations (BSs) is costly as BSs are quite expensive to build and even to maintain. Moreover, BSs consume a significant portion of energy in cellular networks, estimated around 60-80% [4] of the whole network energy consumption as shown in Figure 1-2. A typical mobile phone network in the United Kingdom may consume approximately 40 MW in an hour [6], even excluding the power consumed by the user's handsets. For example, a standard third generation (3G) base station to produce 40W of output radio

frequency (RF) power require 500W of input power, with 12,000 BSs it consumes more than 50GWh of input power every year in a network [6].

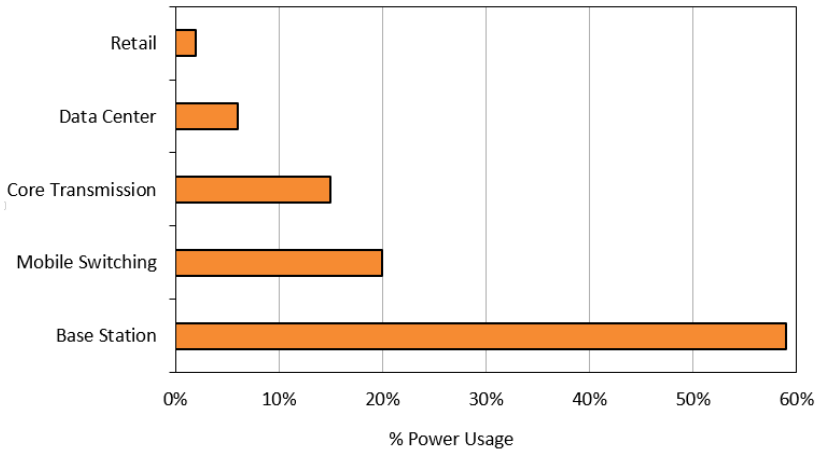

Figure 1-2. Cellular network power consumption [5].

The energy consumption problem in the Information and Communications Technology (ICT) sector has become crucial during the past few years. As a result, governments and industries have recently shown keen concerns on the critical issues related to the energy efficiency in the ICT sector. Among the energy consuming industries, the ICT industry take 2% of global total CO_2 emissions, which is expected to double by 2020 [7]. Also the cellular mobile industry itself is a contributor of CO_2 emission through network operations and mobile equipment due to its tremendous growth and massive involvement in almost every sector. For example, cellular networks are estimated to be responsible for 0.5% of world-wide electrical energy consumption [8]. As most of the energy produced today is still generated from non-renewable energy sources, networks are correspondingly responsible for a significant amount of CO_2 emissions, which is estimated at between 0.5% and 1% of the entire world carbon footprint [8]. The increasing number of wireless devices such as smartphones and high-end wireless devices including tablets, laptops, and M2M nodes and gaming consoles is spawning a raft of data intensive

applications that are accessing mobile networks worldwide. This is one of the primary drivers behind the tremendous growth in data traffic. According to Cisco [9], the global mobile data traffic grew 74% in 2015 and is expected to reach 30.6 Exabyte (EB) per month by 2020, which is an 8-fold increase over 2015 as evidenced in Figure 1-3.

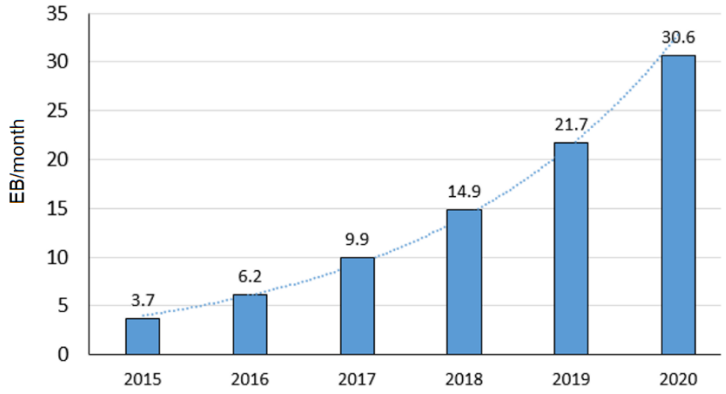

Figure 1-3. Mobile data traffic forecast in Exabyte's per month from Cisco VNI mobile [9].

This figure also reveals that the mobile data traffic will continue predicting the future with certainty at a compound annual growth rate (CAGR) of 53% where:

$$CAGR = [\left(\frac{End\ value}{Starting\ value}\right)^{\frac{1}{number\ of\ years}} - 1]x100$$

A dense deployment of BSs is required to accommodate this massive traffic demand, which correspondingly increases the overall network energy consumption. For example, there are already more than five million BS sites deployed worldwide, with the number expected to grow to more than 11 million by 2020 [10]. The problem is further aggravated by considering financial constraints from the operator's point of view, as higher capacity and better Quality of Service (QoS) come at the cost of higher capital expenditure (CAPEX) and operational expenditure (OPEX). The energy efficiency is a

strategic priority for the operators globally with spending's around $15 billion on energy use annually. Since the BS is the most dominant energy consuming equipment estimated around 60% of the whole network energy consumption, a large electricity bill results from the huge energy consumption of the BS. Mobile communication thus contributes a significant proportion of the total energy consumed by the information technology industry, which is growing significantly year on year. Therefore new methods targeting the reduction of energy consumption by the network entities is a critical requirement for 5G.

It is envisaged that 5G will support a plethora of mobile application like mobile gaming, health monitoring and augmented reality, which are computation intensive and drains a lot of battery when executed in the mobile device itself. The mobile devices can hardly cope due to limitations in terms of battery life, storage, memory and processing power. Extended battery life is also one of the key requirements by mobile phone users as compared to memory, storage and display size, as such there is a need for improvement of energy efficiency in mobile devices as well. This makes energy efficiency in next generation cellular networks a very timely research direction. New techniques are required that can save energy in the mobile device and the mobile network side.

Cloud Radio Access Networks (C-RAN) has been recently proposed by China Mobile in 2010 [11] as a promising solution architecture for reducing energy usage within the cellular networks by performing baseband processing at a centralised cloud computing infrastructure. In C-RAN, a BS is divided into three parts: (i) a remote radio head (RRH) for performing only lower layer RF functions, (ii) high bandwidth fibre fronthaul which connects the RRHs to the

BS cloud and (iii) virtual baseband unit (vBBU) running on virtualised multi-core general purpose processors (GPPs) in the cloud for performing digital baseband processing. The RRH is less intelligent compared to traditional BS's and it is much simpler with lower cost, low failure rate and lower downtime than macro BS. The GPPs consume less power compared to the traditional baseband units (BBUs) used in BS's, are affordable and programmable with the ability to process any signal from any RRH since they are software defined and not hardware dedicated. The GPPs replace digital signal processors (DSPs), field programmable gate arrays (FPGA) and application specific integrated circuits (ASIC) which are vendor locked in and are costly. In C-RAN, cooling, housing and baseband resources are shared in a virtualised cloud infrastructure as such incurring minimal CAPEX/OPEX. As such, HetNets in C-RAN are envisaged as a road towards 5G for improving the energy efficiency.

1.1.1 Main Sources of Energy Consumption in Cellular Networks

According to Nokia Siemens Networks, the energy expenditure within the radio access network (RAN) is shown in Figure 1-4, which interestingly shows only 15% of the energy, is used for forwarding bits [12]. The diagram shows that of the 140TWh produced from the power plant, only 21TWh is used for transmitting bits within the BS. The arrows show the losses in energy. This shows that 85% of the energy is not used for revenue generation but dissipated, since most energy is expended on fans and cooling systems, heating and lighting, uninterruptable and other power supplies, and in running idle resources [13].

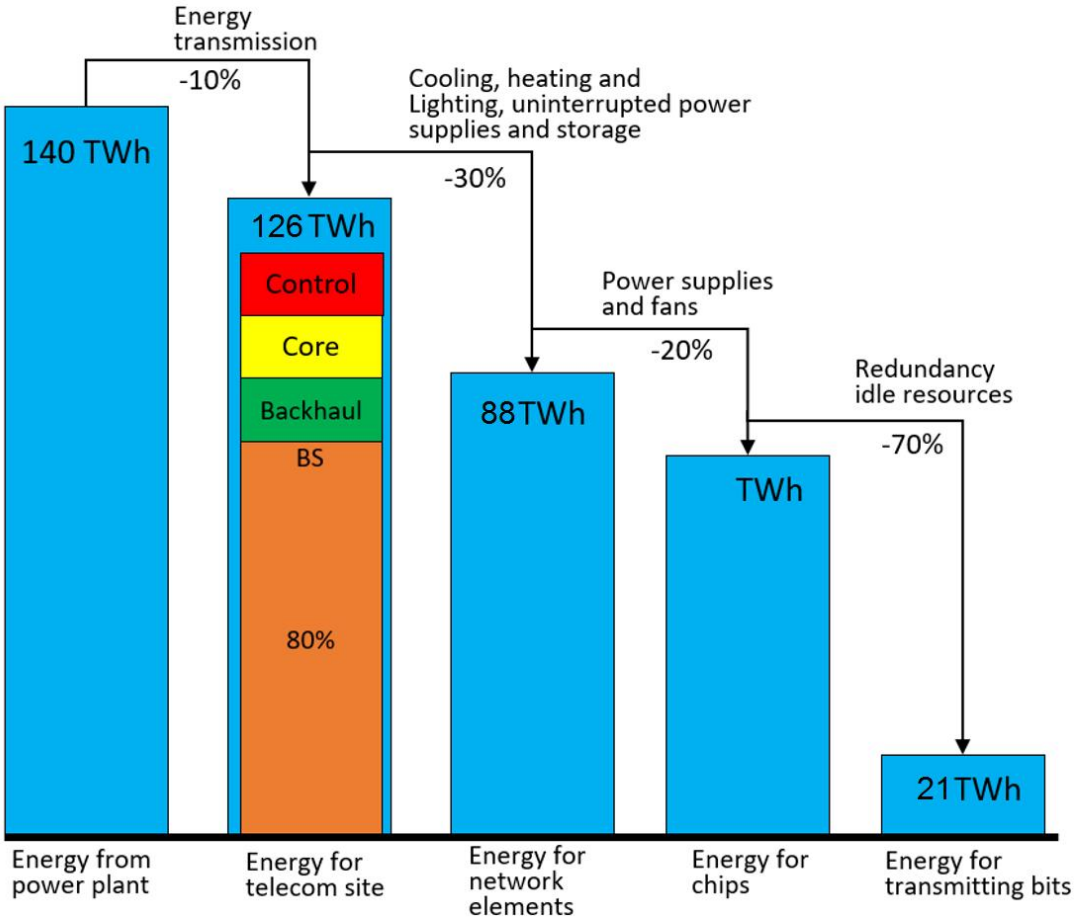

Figure 1-4. Where energy is consumed in a network [12].

Since only a small percentage of energy within RAN is utilised to transfer bits, there is are large potentials and opportunities to reduce these secondary usages of energy to substantially improve the overall energy efficiency of a network. As a result, BS efficiency is one of the key areas in which energy efficiency is ripe for improvement.

In a typical BS, a large amount of power is consumed by the PA and the BBU as shown in Figure 1-5 for both macro and small cells. The energy consumption of BBUs is getting more and more dominant in small cells due to gradual shrinking of cell size and the growing complexity of signal processing

[7]. Hence, it is crucial to optimise the energy efficiency in both the radio front end and BBU servers.

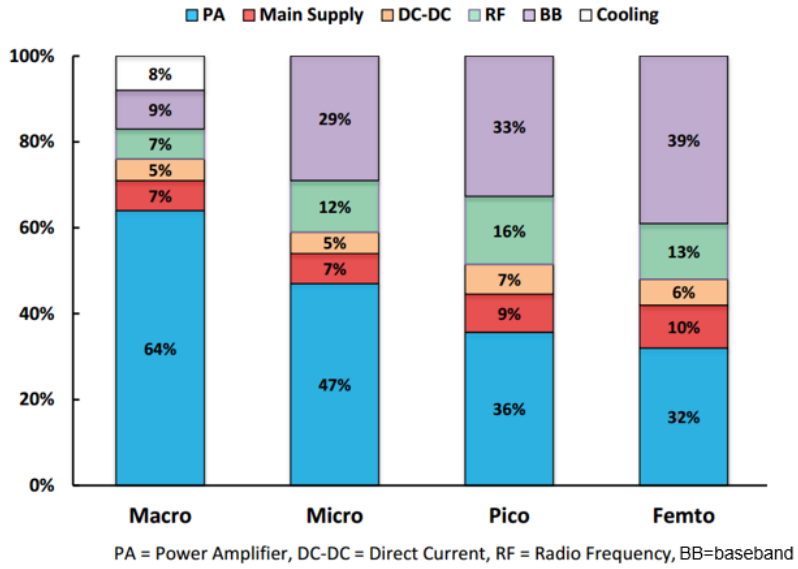

Figure 1-5. BS power consumption for different cell sizes[11]

1.1.2 Cellular Traffic Analysis and Opportunities

It is important to understand the dynamics of cellular traffic in order to find any energy efficient solution. A lot of studies have been carried out for characterizing the traffic generation of aggregate network as well as of individual BS [14]. Lately, several network operators have come forward to disclose their normalised daily traffic patterns [15] and the studies show that across a day, there are significant traffic variations in both temporal and spatial domains. It has also been shown that the peak-time traffic load is much higher (~20 times) than that at off-peak times [15]. Traffic load in this thesis is defined as the ratio of the instant number of active users in the BS or network to the maximum number of users that the BS or network can handle and it is in the range [0,1] or [0%, 100%]. In this thesis, the word ‘traffic load’ and ‘normalised

traffic load' are used interchangeably. The traffic dynamics in both temporal and spatial domains will be broadly described in the following subsections.

1.1.2.1 Temporal Traffic Diversity

Real network data from an anonymous mobile network operator in a metropolitan urban area analysed in [16] was considered to understand the traffic dynamics. The graph in Figure 1-6 shows the normalised traffic trace averaged with a resolution of 30min from five BSs during one week.

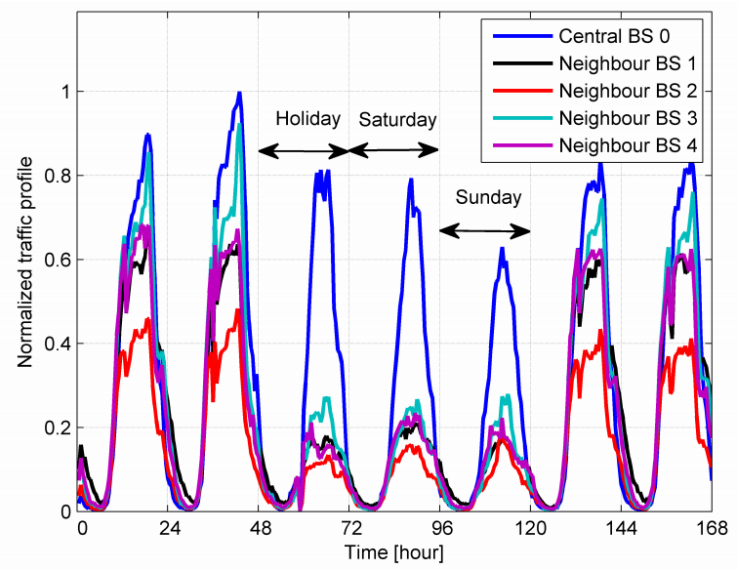

Figure 1-6. Time and spatial domain cellular traffic dynamics over one week [16].

It can be observed that the traffic in each cell is a cyclic variation and it can be noticed that the traffic during daytime (11am – 9pm) or peak periods is much higher than that at night times (10pm – 9am). The temporal variation also depends on the natural life styles and locations. For example, the business areas may be heavily loaded during daytimes but only lightly loaded at night times. Additionally, the traffic profile during weekends/holidays, even during peak hours, is much lower than that of a normal weekday. So, weekdays and weekends appear to show distinct trends as shown in Figure 1-6. Some cells

will always have traffic loads much lower than the network capacity for a large fraction of time in a day (8~10 hours), which indicates the underutilisation of each BS in the temporal domain. Indeed, this will result in network wide energy inefficiency at BSs.

1.1.2.2 Spatial Traffic Diversity

The traffic in different regions can be very different due to user mobility, the behaviour and activities of mobile subscribers on the temporal scale and demands for data and video applications. In the city centre (hot spots), traffic demands are very high during the day time whereas the traffic demand may not be the same in residential areas. Similarly, the traffic demands in cells near railway stations are very high compared to other areas. So, the traffic loads can vary greatly even in neighbouring cells. The normalised traffic over three consecutive weeks taken from three different cells of the same network was investigated in [17] and the results are plotted in Figure 2-11.

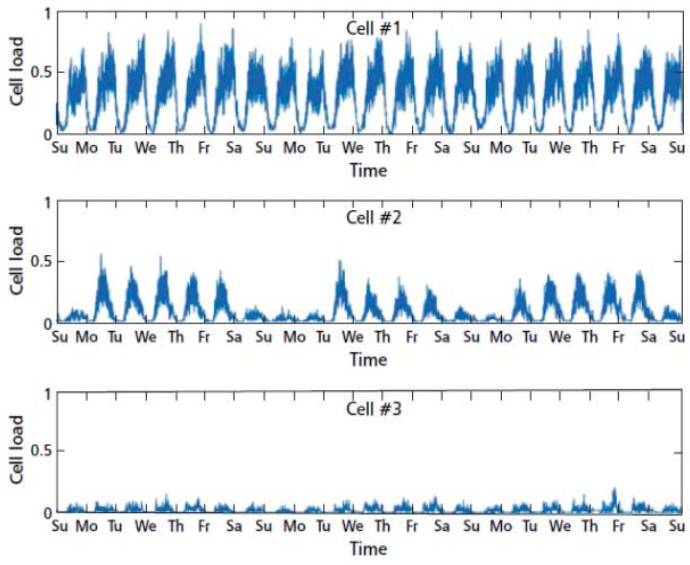

Figure 1-7. Normalised load of three different cell over three weeks showing high load (Cell#1), varying load (Cell#2) and low load (Cell#3) [17]

It shows clear evidence of the spatial diversity as the traffic generation is much lower in cell#3 compared to the other two.

In summary, the traffic in a cellular network is quite diverse over time and space. Such strong temporal and spatial traffic diversity indicate the underutilisation of BS resulting in both system and network-wide energy inefficiencies at BS. However, the traffic dynamics can provide significant opportunities for energy savings. For example, if the traffic variation can be traced and the resource allocation strategy for individual or the whole network is adopted accordingly, a significant amount of energy could be saved. From the above discussion, it can be concluded that the variation of traffic density in cellular networks in both the temporal and spatial domains shows significant under-utilisation of system capacity given the network being designed based on the peak-traffic scenarios.

1.2 Problem Statement and Motivation

Demand for mobile services is currently exploding, posing major challenges to mobile operators in supporting these high capacity requirements and improving QoS. In order to meet throughput demands, the 3rd generation partnership project (3GPP) introduced HetNets in long-term evolution-advanced (LTE-A) [3] where macro cells are overlaid by small cells. However, a massive deployment of small BSs result in considerable increase in HetNet energy consumption.

Traditional distributed RAN architectures consume a significant amount of energy and waste a lot of computing power as the radio front ends are always kept on and also the BBU processing servers are underutilised as they are not

shared but serve each individual cell [7] [8]. Traditionally, BSs have been preconfigured to provide peak capacities as such, the BS equipment's are kept active even when there are no users within the BS which leads to unnecessary energy consumptions. Nevertheless, even in peak hours, 90% of the data traffic is carried by only 40% of the cells [9] which gives room for energy improvement as the other cells with low utilisation can be switched off. The mobile traffic varies significantly, irrespective of both the time of day or traffic profile and is rarely at its peak in practical scenarios. This means traffic load in a cellular network changes gradually in a time-geometry pattern called the 'tidal effect' which is the fluctuation of traffic load in the BS due to the dramatic subscriber density increase in both business and residential areas.

Traditional BSs are also very expensive to buy and maintain hence have high OPEX and CAPEX. Also as the BS processing capacity is only used for its own coverage rather than being shared in a large geographical area, this amounts to huge BS underutilisation. During the evening BSs in residential areas are over-subscribed whereas BSs in business areas stay under-subscribed. However, these undersubscribed BSs still consume a significant amount of energy even when they are not necessarily required to be kept active. Therefore, it is imperative to find innovative solutions that reduce energy consumption by adapting the required network resources to the traffic demands and locations.

Energy consumption is not only experienced in the BSs, but also in the mobile devices which may run several computing intensive applications that degrade battery life. Therefore, new solutions are required that can improve the battery life in mobile devices like offloading mobile applications to a public cloud.

Moreover, with billions of devices connected to the 5G network, traffic congestion will be an issue. As such there is a need for efficient call admission control (CAC) mechanisms that can avoid traffic congestion which can reduce blocking probability within the network and improve BS resource utilisation.

1.3 Aim and objectives

The main aim of this PhD project is to design an energy efficient resource management framework for 5G HetNet C-RAN based system which aims at minimising energy consumption and at the same time maximising resource utilisation while maintaining QoS. The following objectives have been set to meet the defined aim of this study.

1. Develop a suitable energy model for 5G C-RAN system that accurately models the total energy consumed in the proposed C-RAN architecture and considers various aspects like network traffic load, RRH, virtualised BSs, fronthaul and the BS cloud.
2. Design a novel energy efficient resource management framework for 5G RAN that incorporates BS sleeping and baseband workload consolidation techniques.
3. Design an energy efficient resource management framework for saving energy in mobile devices within 5G C-RAN that incorporates breaking down a mobile application into tasks that can be processed in the BS cloud in virtualised MEC servers.
4. Design a novel CAC framework for 5G C-RANs that leverages the proposed resource management framework to avoid traffic congestion.
5. To evaluate the performance of the proposed framework for different scenarios.

1.4 Contributed Work and Achievements

The main contributions of this thesis include:

1. A novel computational-resource-aware energy consumption model for 5G C-RAN is derived. The proposed model considers the separation of the RF function called RRH, fronthaul and BS cloud power consumption. Traditional BS power consumption models cannot be used for our scenario, since in C-RAN, baseband processing power, cooling and housing are shared in the BS cloud.
2. A novel dynamic centralised BS switch-off mechanism at the radio side of 5G C-RAN while insuring minimal dropping probability is proposed. In the proposed mechanism, only the small cells are dynamically switched off based on a proposed utility function whereas the macro cells are always kept active for maintaining coverage. Traffic of the low utility small cells will then be offloaded to overlaying macro-BS or overlapping small cell if there are enough resources. The utility function considers various factors including the total rate of served users, the small cell traffic load, the received interference signal strength from the nearby cells and power consumption of the BSs.
3. A new energy-efficient BBU workload consolidation resource management algorithm at the BS cloud of 5G C-RAN is proposed. This algorithm uses cloud computing based baseband workload consolidation technique for minimising energy consumption. Here physical BBUs are replaced with virtualised GPPs and the rationale is to reduce the number of GPPs used by migrating vBBUs from less loaded physical GPPs to other GPPs and switching off idle GPPs during low traffic periods. The GPPs minimisation

problem is solved using bin-packing algorithms like Next Fit (NF), First Fit (FF), First Fit Decreasing (FFD) and Full Bin Packing (FBP) algorithms. Moreover novel Simulated Annealing (SA) and Genetic Algorithm (GA) based algorithms are also proposed to minimise BS cloud energy consumption.

4. A novel fuzzy logic based CAC framework is proposed for 5G C-RANs that also considers the pre-emption of call with the aim to improve GPP utilisation and reduce call blocking probability and boost operator revenue. During congestion a technique called cloud bursting is adopted where delay tolerant non real-time (NRT) low priority connections are pre-empted and outsourced to a public C-RAN cloud with a pricing penalty to accommodate the real-time (RT) connections in the private operator C-RAN.
5. An energy efficient framework for saving energy in mobile devices within 5G C-RAN using mobile edge computing (MEC) is also proposed. The proposed technique provides a paradigm shift to 5G as cloud computing capabilities are provided in close proximity to mobile devices and not at distant locations. In the proposed framework, an application from the mobile device is partitioned into modules/tasks which are either executed on the local device or offloaded and executed in the BS cloud while considering the transmission power, application size, QoS, state of charge (SoC) of the mobile battery and mobile device central processing unit (CPU) load.

1.4.1 Published Work

The following materials have been published taking material from the thesis which form the core contribution of this thesis.

Published Journals

1. **T. Sigwele**, A. S. Alam, P. Pillai, and Y. F. Hu, *"Energy-efficient cloud radio access networks by cloud based workload consolidation for 5G"*, Journal of Network and Computer Applications (JNCA), vol. 78, pp. 1-8, 2017**. (3.5 Impact Factor)**

Submitted Journals

1. **T. Sigwele**, A. S. Alam, P. Pillai, and Y. F. Hu, *"Energy-Efficient 5G Cloud RAN with Virtual BBU Server Consolidation and Base Station Sleeping"*, IEEE Transaction on Cloud Computing, 2017**.**
2. **T. Sigwele**, A. S. Alam, P. Pillai, and Y. F. Hu, *"Fuzzy-Logic Based Call Admission Control in 5G Cloud Radio Access Networks with Pre-emption"*, EURASIP Journal on Wireless Communications and Networking, 2017**.**

Published Conference Papers

1. **T. Sigwele**, P. Pillai, and Y. F. Hu, " Saving Energy in Mobile Devices Using Mobile Device Cloudlet in Mobile Edge Computing for 5G," in *IEEE International Conference on Green Computing and Communications (GreenCom-2017),* Exeter, UK, 2017.

2. **T. Sigwele**, A. S. Alam, P. Pillai, and Y. F. Hu, "On Energy Minimisation of Heterogeneous Cloud Radio Access Networks," in *International Conference on Wireless and Satellite Systems (Wisats),* Cardiff, UK, 2016.

3. **T. Sigwele**, A. S. Alam, P. Pillai, and Y. F. Hu, "Evaluating Energy Efficient Cloud Radio Access Networks for 5G," in *The 11th IEEE International Conference on Green Computing and Communication (Greencom)*, Sydney, Australia, pp. 362-367, 2015.
4. **T. Sigwele**, P. Pillai, and Y. F. Hu, "iTREE: Intelligent Traffic and Resource Elastic Energy Scheme for Cloud-RAN," in *3rd IEEE International Conference on Future Internet of Things and Cloud (FiCloud),* Rome, Italy, pp. 282-288, 2015.
5. **T. Sigwele**, P. Pillai, and Y. F. Hu, "Elastic Call Admission Control Using Fuzzy Logic in Virtualised Cloud Radio Base Stations," in 7th *Wireless and Satellite Systems (WiSATS2015)*, Bradford, UK, 2015. **(BEST PAPER AWARD)**
6. **T. Sigwele**, P. Pillai, and Y. F. Hu, "Call Admission Control in Cloud Radio Access Networks," in *2nd IEEE International Conference on Future Internet of Things and Cloud (FiCloud)*, Barcelona, Spain, pp. 31-36, 2014.

Submitted Conference Papers

1. **T. Sigwele**, P. Pillai, A. Sangodoyin and Y. F. Hu, " Security Aware Virtual Base Station Placement in 5G Cloud Radio Access Networks," in *International Conference on Wireless and Satellite Systems (Wisats)*, Oxford, UK, 2017.

Book chapter

1. **T. Sigwele**, P. Pillai, and Y.F. Hu, "Elastic Call Admission Control Using Fuzzy Logic in Virtualised Cloud Radio Base Stations," in *Wireless and Satellite Systems*. vol. 154, P. Pillai, Y. F. Hu, I. Otung, and G. Giambene, Eds., ed: Springer International Publishing, pp. 359-372, 2015.

1.5 Structure of the Thesis

This thesis is divided into 6 chapters. Chapter 1 presents a brief introduction of the research topic, states the overarching aim and objectives of the work and also highlights the problem statement. Chapter 2 begins by presenting an overview of RANs including generations and challenges of today's RANs and then introduces the concepts and motivation of C-RAN. A literature review of energy efficiency in C-RAN and MEC is discussed in Chapter 3. Chapter 4 presents the detailed description of the proposed energy efficient 5G C-RAN framework. The system model and problem formulation are presented here along with the proposed energy consumption model. The proposed BS sleeping scheme, the baseband processing workload consolidation framework, the proposed fuzzy based CAC with pre-emption and saving energy on the mobile device using MEC are presented in this chapter. Chapter 5 then discusses the simulation platforms used and the various simulation scenarios. This chapter also presents the results obtained for these different scenarios to critically evaluate the performance for the proposed framework. Finally, Chapter 6 presents the thesis conclusion which summarises the contributions made by this research and also discusses some recommendations for future work and development.

2 Radio Access Networks

This chapter presents an overview of the evolution of the various cellular wireless generations from first generation (1G) to fourth generation (4G) and presents the challenges faced by today's RANs. The next generation cloud computing based RANs such as C-RAN and MEC will be comprehensively discussed including the motivation, architecture for C-RAN, evolution of C-RAN and MEC.

2.1 Radio Access Networks Overview

Wireless communication has become a ubiquitous part of modern life, from global cellular communication systems to local and even personal area networks. Even to the most casual observer, it is apparent that a veritable revolution in telecommunications has taken place within recent years. The use of wireless communications has expanded dramatically globally, as more and more users are using data applications [18]. The past years have experienced a phenomenal growth in the wireless industry, both in terms of mobile technology and its subscribers. There has been a significant shift from fixed to mobile cellular telephony, especially since the turn of the century. The cellular concept was a major breakthrough in solving the problem of spectral congestion and user capacity [19]. It is a system-level idea which calls for replacing a single, high power transmitter (covering a large area) with many low power transmitters covering small geographic areas called cells that are represented as a hexagon. Each cell is served by a BS providing coverage to only a small portion of the service area.

Figure 2-1 shows an evolution of wireless access networks from 1G to 5G with Wireless Local Area Networks (WLAN) internetworking with cellular networks in 5G networks. The cellular wireless generation generally refers to a change in the fundamental nature of the service, non-backwards compatible transmission technology, and new frequency bands. New generations have appeared in every ten years.

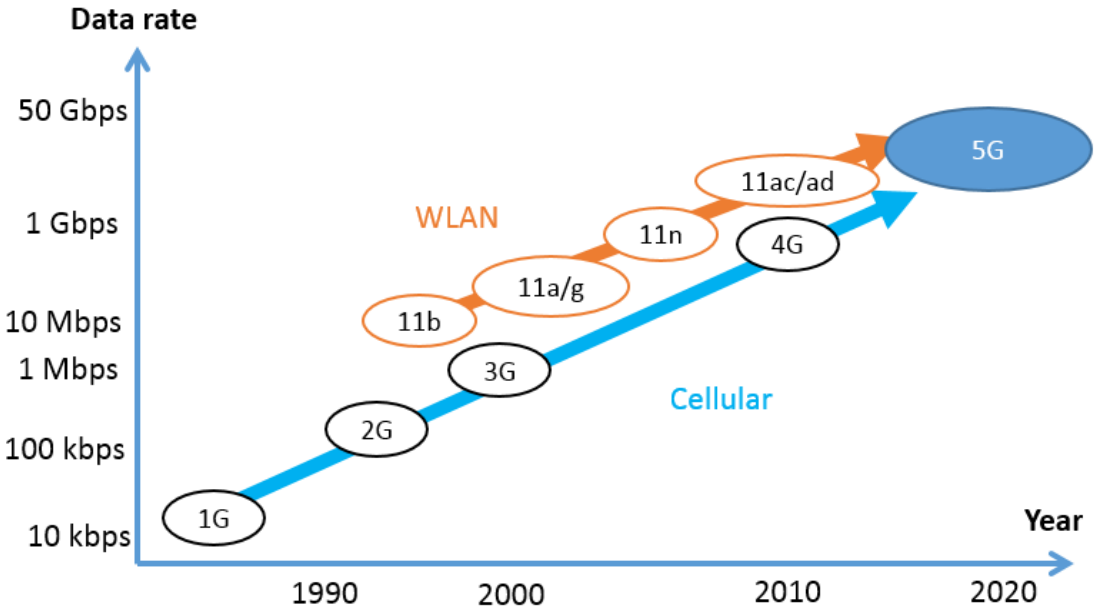

Figure 2-1. Evolution of wireless communication technology [20].

2.2 Generations of Radio Access Networks

2.2.1 First Generation (1G) Mobile Systems

1G of mobile wireless communication were based on the analog transmission system for speech services [21]. 1G was first introduced in 1979 by Nippon Telephone and Telegraph (NTT) in Tokyo, Japan as a significant leap in mobile communication, especially in terms of capacity and mobility. Two years later, the cellular epoch reached Europe. In the United States, the Advanced

Mobile Phone System (AMPS) was launched in 1982. The system was allocated a 40MHz bandwidth within the 800-900MHz frequency range by the Federal Communications Commission (FCC). AMPS offered 832 channels, with a data rate of 10Kbps. Although omnidirectional antennas were used in the earlier AMPS implementation, it was realised that using directional antennas would yield better frequency reuse. Directional antennas yield better frequency reuse since a BS can have many sectors using various directional antennas at different frequencies. Total Access Communication System (TACS) was deployed in the United Kingdom in 1983. 1G cellular system had very low data rates of 2.4Kbps and were analogue based allowing only voice calls. The drawbacks of 1G telephony were [22]:

- Poor voice quality as analogue signals are easily affected by interference.
- Poor battery life.
- There was no security of data as analog signals did not allow advanced encryption methods. Hence, anybody could listen to the conversation easily by simple techniques.
- Limited capacity and poor handoff reliability.

2.2.2 Second Generation (2G) Mobile Systems (Digital)

In the end of 1980s, the Second Generation (2G) mobile systems were announced, and were launched on the Global System for Mobile Communications (GSM) standard which used audio quality digital modulation to provide voice and limited data services [21]. Compared to 1G systems, 2G systems used digital multiple access technology, such as Time Division

Multiple Access (TDMA) and Code Division Multiple Access (CDMA). Three primary advantages of 2G networks over 1G were that voice conversations were digitally encrypted; 2G systems were significantly more efficient on the spectrum over their predecessors; and data services for mobiles were introduced in 2G like Short Messaging Service (SMS), fax and paging. In the late 1990s, 2.5G was introduced which used the General Packet Radio Service (GPRS) standard with improved data rates of 64-144Kbps [23]. 2.5G was developed in between its predecessor, 2G, and its successor, third generation (3G). Later on, Enhanced Data Rates for GSM Evolution (EDGE), also dubbed 2.75G, was launched as a mobile technology that allows improved data transmission rates as a backward-compatible extension of GSM.

2.2.3 Third Generation (3G) Mobile Systems

3G is the 3rd generation of mobile telecommunications. 3G systems support services that provide an information transfer rate of at least 2Mbps. It uses wide band wireless network with which clarity is increased and therefore satisfying the International Mobile Telecommunications-2000 (IMT-2000) specifications by the International Telecommunication Union (ITU). Universal Mobile Telecommunications System (UMTS) is a European 3G standard developed by 3GPP. CDMA2000 is the American 3G standard developed by 3GPP2 in 2001. 3G finds applications in wireless voice telephony, mobile internet access, fixed wireless internet access, video calls and mobile television. Later 3G releases provide mobile broadband access of several Mbps to smartphones and mobile modems in laptop computers.

2.2.4 Fourth Generation (4G) Mobile Systems

4G is the successor of 3G and it aims to provide mobile broadband internet access. Some of its possible applications include amended mobile web access, Internet Protocol (IP) technology gaming services, high definition mobile television, three dimension (3D) television, video conferencing and cloud computing. Two 4G candidate systems are commercially deployed: the Mobile Worldwide Interoperability for Microwave Access (WiMAX) standard and Long Term Evolution (LTE) standard [21]. LTE uses Orthogonal Frequency Division Multiplexing (OFDM) and OFDM access, which divides a channel usually 5, 10 or 20MHz wide into smaller sub channels or subcarriers each 15 KHz wide. Each is modulated with part of the data with one of several modulation schemes like Quadrature Amplitude Modulation (QAM), 16QAM and 64QAM. LTE also defines MIMO operation that uses several transmitter-receiver-antennas. The data stream is divided between the antennas to boost speed and to make the link more reliable. Using OFDM and MIMO let LTE achieve data rates of up to 100 Mbps downstream and 50 Mbps upstream under the best conditions. LTE Advanced (LTE-A) is an improvement of LTE and is the candidate for IMT-Advanced standard formally submitted by the 3GPP organisation to ITU in the fall 2009. The target of 3GPP LTE-A is to reach and surpass the ITU requirements. The peak downlink data rate for LTE-A is 1Gbps and peak uplink data rate is 500Mbps. 4G traffic exceeded 3G traffic for the first time in 2015 according to [9]. The Institute of Electrical and Electronics Engineers (IEEE) 802.16m (WiMAX) is under development, with the objective to fulfil the IMT-Advanced criteria of 1Gbps for stationary reception and 100Mbps for mobile reception.

2.2.5 Fifth Generation (5G) Mobile Systems

5G denotes the next major phase of mobile telecommunications standards beyond the 4G/IMT-Advanced standards [24]. At present, standardisation activities are being carried out in this area. It is expected that the 5G mobile system will be an all-IP based model for wireless and mobile networks interoperability. Below is a summary list of 5G requirements [25] from various 5G bodies such as 5G Infrastructure Associate (5G IA), 5G Forum, 5G Mobile Communication Promotion Forum (5GMF) and 5G Infrastructure Public Private Partnership (5G-PPP).

1. Higher system capacity: The target is set to achieve a 1000-fold system capacity per km^2.
2. High data rates: 5G should target to provide multi-Gbps transmission rates up to 20Gbps along with more uniform Quality of user Experience (QoE) compared to 4G.
3. Supports massive connectivity: 5G should support up to 10-100 times more connected devices which can be up to 100 billion devices.
4. Zero latency: 5G has to provide not only higher data rate, but also a negligible user plane latency of less than 1ms over the RAN, a large leap from LTE's 5ms.
5. Higher energy efficiency: 5G target to save energy in both the RAN and the terminals with up 90% reduction in network energy consumption and also up to 10 years battery life for low power machine type device.
6. Mobility: Target maximum speed of up to 500km/h.
7. Reliability: 99.999% reliability and 100% coverage

8. Network flexibility: Multi RAT system, Network Function virtualisation (NFV) and Software Defined Networks (SDN).
9. Flexible Spectrum: Efficient integration of existing cellular bands and new spectrum bands over a wide range of frequency bands from the current frequency bands below 6GHz to above 6GHz (30GHz-300GHz) mmW bands incorporating carrier aggregation, operation in unlicensed spectrum bands and cognitive radio.
10. High density networks: Advanced small cells and Self-Organizing Networks (SON).

The NGMNA feels that 5G should be rolled out by 2020 to meet business and consumer demands [26]. In addition to providing simply faster speeds, it is predicted that 5G networks also will need to meet new use cases, such as [27]:

1. Enhanced mobile broadband (eMBB): Virtual and augmented reality, 3D video, holographic presence.
2. Massive machine type communication (mMTC) for IoT devices: e-health, millions of sensors connected, smart home.
3. Ultra-reliable and low latency communications (uRLLC): remote surgery, vehicle to everything communication, self-driving cars, drone delivery, smart manufacturing.

2.3 Challenges of today's RAN's

Currently the RANs of today are facing the following challenges:

1. Increased capacity needs
2. High power consumption

3. Low BS utilisation
4. Increased CAPEX and OPEX
5. Increased interference

2.3.1 Increasing Capacity needs with Flat Average Revenue per User (ARPU)

Figure 2-2 shows how the end user's mobile data consumption accelerates as the datarates of mobile broadband network develops significantly with the introduction of air-interface standards such as 3G and 4G. It is envisaged that with the deployment of 5G in the next few years, the number of users who access mobile broadband will triple [9].

Figure 2-2. Global mobile traffic by connection type [9].

These findings mirror the fact that the expanding bandwidth of wireless broadband triggers the expansion in mobile traffic, since the mobile users can utilise an assortment of high-bandwidth services, like video-based applications. This pattern poses a serious challenge to the next generation RAN. The global mobile traffic increases 66-fold with a CAGR of 131%

according to [9]. The peak datarate on the contrary increases with a CAGR of 55% from UMTS to LTE-A. As shown in Figure 2-3, there is a large gap between the mobile network operator revenue and global mobile data traffic because the incremental revenue growth no longer supports the network costs required to cope with the surge in mobile data traffic. In addition, additional reductions in data costs are necessary to maintain profitability for mobile operators. New infrastructure technologies need to be developed to further enhance the performance of LTE/LTE-A in order to fill this gap.

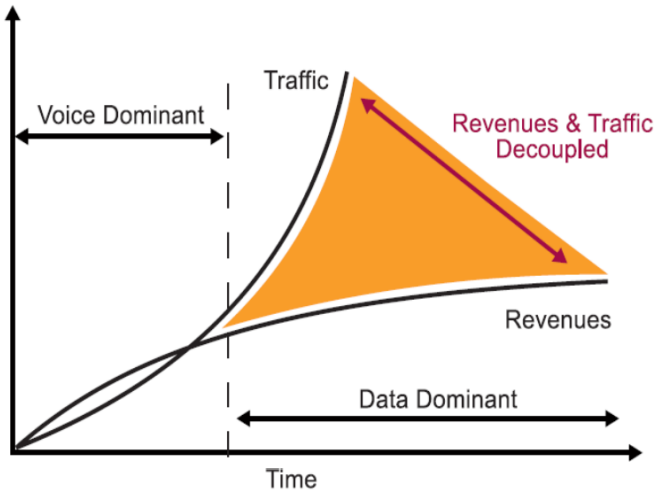

Figure 2-3. Growing gap between the mobile network operator revenue and global mobile data traffic [28].

On the other hand, the income of mobile operators is not increasing at an indistinguishable pace as the network capacity they provide. Hence, the Average Revenue per User (ARPU) is falling whereas the data volume grows quickly. To confront the slow growth in revenue, operators are compelled to continually hold down expenses. That implies that mobile operators must find a high-capacity, low cost access network with novel strategies to meet the

development of mobile data traffic while keeping a sound and profitable growth.

2.3.2 High Power Consumption

2.3.2.1 Power Consumption in BSs

The power consumption by telecom networks is constantly rising as the operators increase the number of BSs to offer broadband wireless services in smaller dense cells. For instance, China Mobile has over 1.44 million 4G BSs [11] to give better network coverage and capacity. Subsequently, the total power consumption has also multiplied. The higher power consumption results in critical environmental effects and higher OPEX, both of which are now increasinly unacceptable. Figure 2-4 demonstrates the components of the power consumption of China Mobile [11]. The figure shows that the dominant power consumption of around 72% is from the BSs in the RAN.

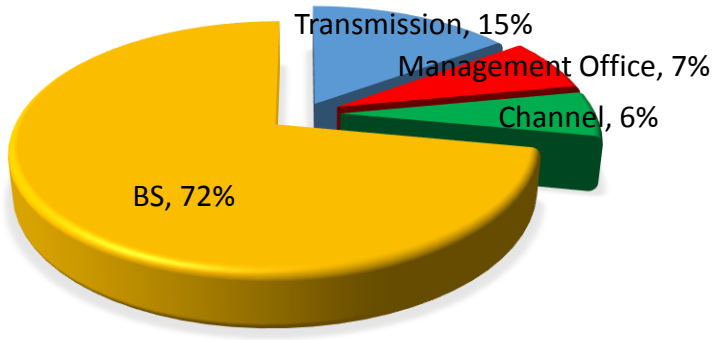

Figure 2-4. Power consumption within RAN [11].

Obviously, the ideal approach to save energy and reduce carbon-dioxide emissions is to reduce the number of BS within the coverage area. Nevertheless, this will results in poor network coverage and lower network

capacity within the RAN coverage area. As a result, operators are looking for new strategies to reduce energy consumption without decreasing the network coverage and capacity. Today, there are a significant number of amendment strategies that help reduce BS power consumption, like the software solutions which save power through switching off selected carriers during off peak hours like midnight, the environmentally friendly energy solutions which offer solar, wind and other renewable energy for BS power supply according to local natural conditions, and the cooling innovations which decrease the energy consumption of the cooling hardware. Nonetheless, these advancements are incremental strategies and cannot address the fundamental issues of power consumption with the number of increasing BSs. A radical transformation in the infrastructure is the key to resolve the power consumption challenge. Virtualised BS like in C-RAN as discussed in Section 2.4 would decrease the number of BS equipment rooms, lessen the air conditioning need, and utilise resource sharing technologies to enhance the BS utilisation rate efficiency under dynamic network load.

2.3.2.2 Power Consumption in Mobile Devices

The power consumption in mobile phones is also important and according to Qualcomm [29], 72% of users rate battery life as the most important feature of a mobile device as compared to screen size, storage, wireless fidelity (WiFi) speed as shown in Figure 2-5. Consumers rank phone battery life as the most important factor in their smartphone buying decision. The most common causes of excessive power consumption include the display, peripherals, processor, audio, RF module as shown in Table 1.

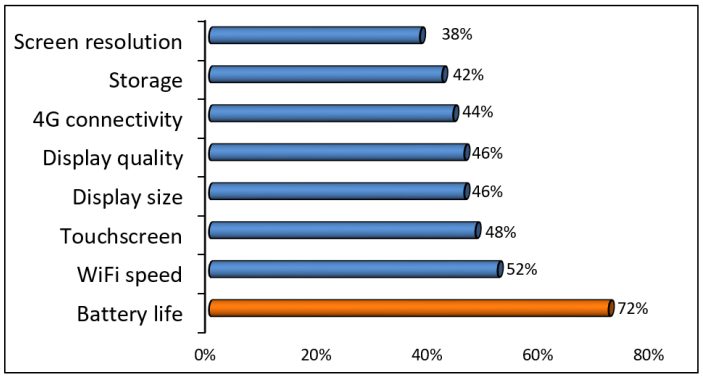

Figure 2-5. Importance of smartphone features among smartphone buyers [29].

Table 1. Power consumption in a mobile device [30].

Function	Power consumption
Display	900mW
Peripherals	1500mW
Processor	1620mW
Audio	400mW
RF	1330mW
Total	**5750mW**

One other feature that consumes battery power is computing intensive applications. With the introduction of 5G in 2020, more diverse applications are developed for mobile devices like smartphones and tablets. These new applications drain the battery 2 to 5 times faster than normal. According to [29], over 70% of these applications are chart toppers in google play and some of these applications can cause unexpected data usage (up to 2.2GB per month). Heavy battery usage in mobile devices is the top reason consumers uninstall applications. Hence it is imperative that new solutions are researched that can reduce the battery usage of the mobile devices but still allow the uptake of these applications.

2.3.3 Dynamic Mobile Network Load and Low BS Utilisation Rate

The dynamic nature of mobile network load has been introduced Section 1.1.2 where the network load moves in the mobile network with a pattern called the tidal effect, as demonstrated in Figure 2-6, where during working hours, the office area BSs are the busiest and in the non-work hours, the residential or entertainment areas are the busiest.

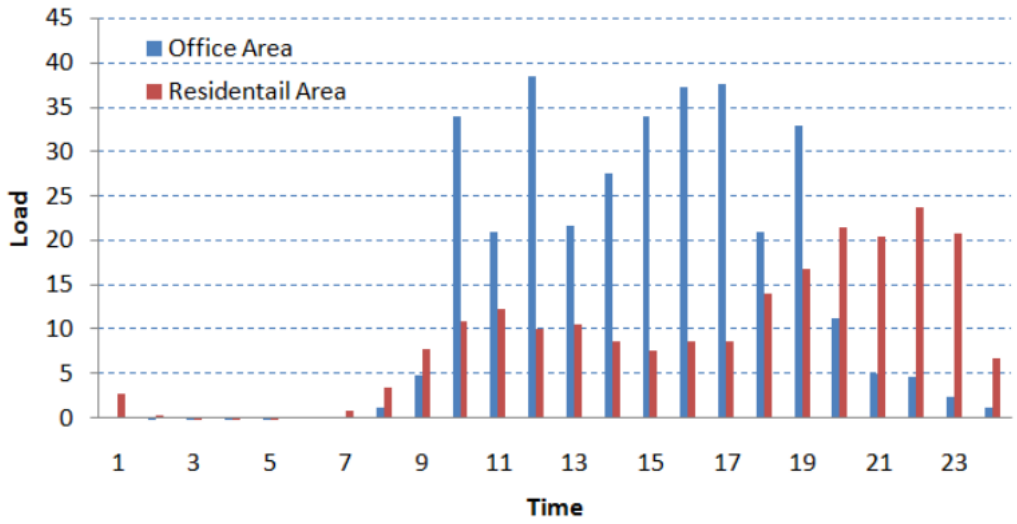

Figure 2-6. Mobile network load in daytime [11].

In the traditional RAN, each BS's baseband processing capability can only be used by the active users within its cell range, causing idle BS in some zones/times and oversubscribed BS in other zones. When subscribers move to other areas, the BS just stays in idle mode with a large portion of its processing power wasted as shown in Figure 2-7. These idle BSs consume almost the same level of energy as they do in busy hours since operators must provide 24/7 coverage. Even worse, the BSs are dimensioned to handle a maximum number of active subscribers in busy hours, as such they are designed to have much more capacity than the average required, which

means that most of the processing capacity is wasted in non-busy time. A way to utilise these BS more effectively is sharing the baseband processing and thus the power consumption between different cell areas which is a C-RAN concept described in detail in Section 2.4.

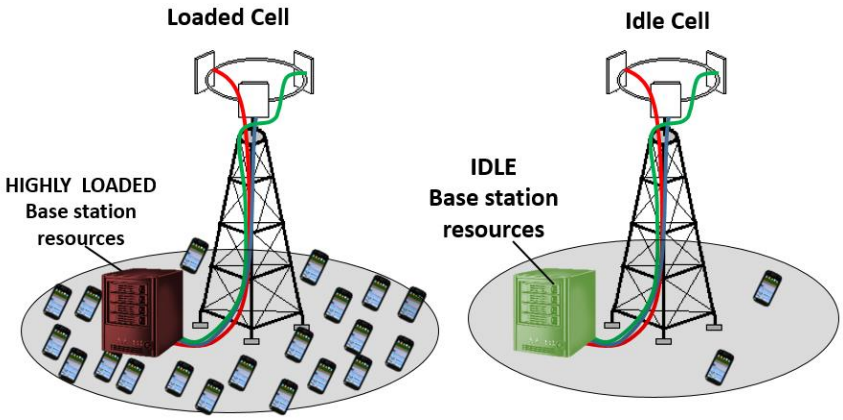

Figure 2-7. Resource wastage in traditional RAN.

2.3.4 Increased Capital and Operational Expenditure (CAPEX/OPEX)

In order for the operators to remain profitable and competitive, they become more and more cautious about the Total Cost of Ownership (TCO) of their network. The TCO, including the CAPEX and the OPEX are shown in Figure 2-8 and Figure 2-9 respectively. The CAPEX is mainly related with network infrastructure building, whereas OPEX is mainly associated with network operation and management. Generally, up to 80% CAPEX of a mobile operator is spent on the RAN. This implies that a greater part of the CAPEX is related to building up cell sites for the RAN. The CAPEX is predominantly spent at the phase of cell site constructions and consists of purchase expenditure (purchases of BS and supplementary equipment, such as power and air conditioning equipment, etc.) and construction expenditures (network

planning, site acquisition, civil works and so on). OPEX has a huge impact in the TCO in network operation and the maintenance phase. OPEX includes the cost of site rental, transmission network rental, operation/maintenance and bills from the power provider.

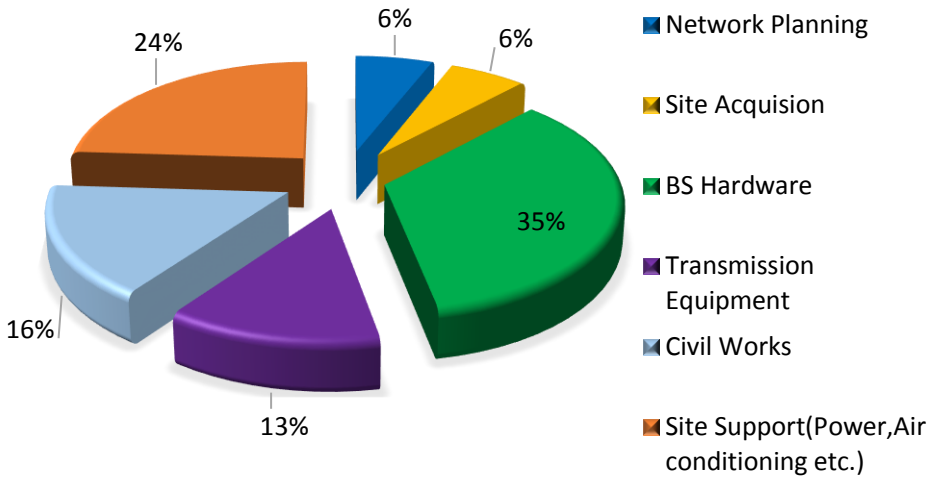

Figure 2-8. CAPEX analysis of cell site [11].

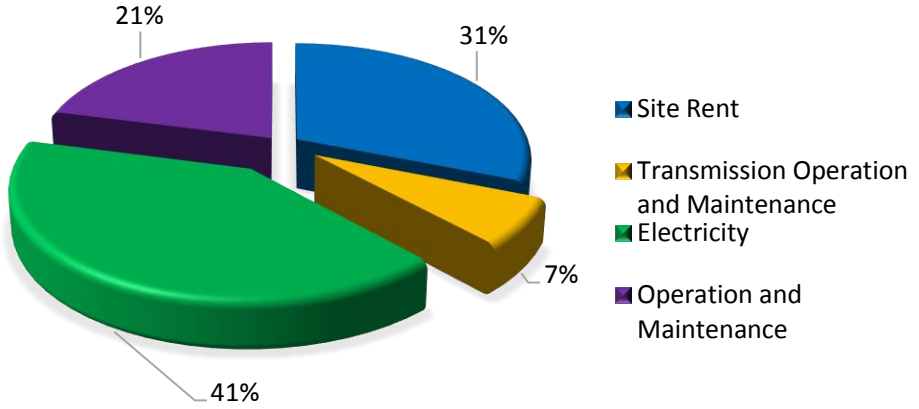

Figure 2-9. OPEX analysis of cell site [11].

2.3.5 Increased interference

Wireless mobile communication systems such as 2G and 3G operate at a frequency reuse factor greater than one. The frequency reuse factor is the rate at which the same frequency can be used in the network. The conventional

LTE system used frequency reuse factor of one as in Release 8. However, this caused greater interference especially at the cell edge, which reduces the SINR for cell-edge users. As such, Fractional frequency Reuse (FFR) has been proposed to mitigate cell edge interference where a cell's bandwidth is partitioned using more bandwidth compared to the conventional LTE system resulting in frequency reuse factor greater than one. The essential downlink and uplink transmission technologies for LTE are OFDM and Single-Carrier Frequency Division Multiple Access (SC-FDMA) respectively. The orthogonality among different sub-carriers eliminates the intra-cell interference. However, since every cell operates on the same frequency band, the inter-cell interference from and to the adjacent cells becomes distinctly unavoidable, which leads to low-throughput performance. How to avoid and eliminate inter-cell interference becomes an important researching subject for future RANs, which is outside the scope of this thesis.

2.4 Cloud Radio Access Networks (C-RAN)

C-RAN, sometimes referred to as centralised-RAN [31-33], super BS [34] or Airvana's onecell [35], is a proposed architecture for future cellular networks where baseband processing resources called BBUs are pooled to a central office, so that they can be shared between BSs. C-RAN was first introduced by China Mobile Research Institute in April 2010 [11]. C-RAN can support 2G, 3G, 4G systems and future wireless communication standards. Whereas traditional cellular networks are built with many stand-alone BSs, C-RAN consolidate BSs to a central cloud called baseband pool or BS cloud. In the following sections, the C-RAN taxonomy as shown in Figure 2-10 will be described.

2.4.1 Evolution of C-RAN

This section describes the traditional and the new evolved RAN architectures which mainly differs in the way the BS functions are located and structured.

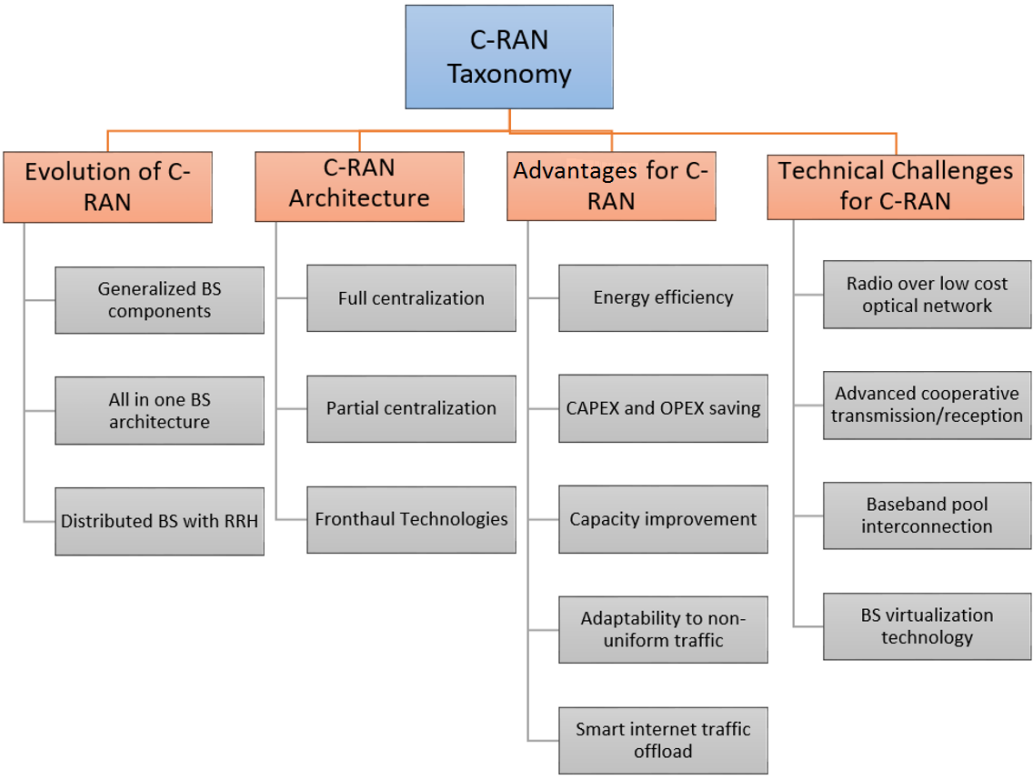

Figure 2-10. C-RAN taxonomy.

2.4.1.1 Generalised BS Components

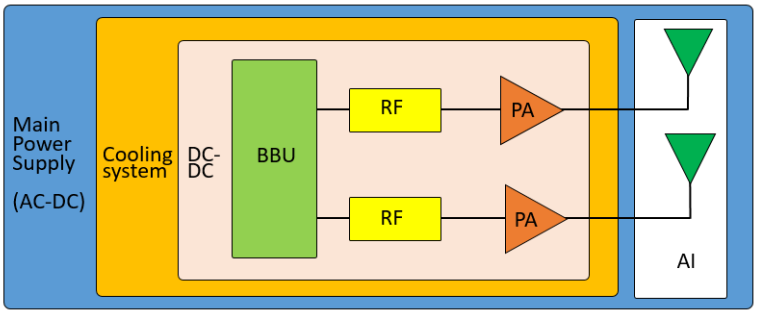

Figure 2-11. A simplified block diagram of a BS with its main power consuming components [36].

A simplified block diagram of a complete BS that can be generalised for all BS types is shown in Figure 2-11. A BS typically consists of different power consuming components and the amount of component energy consumption varies depending on the type of BS. A BS consists of a number of Antenna Interfaces (AI), Power Amplifiers (PA), RF transceiver sections, a BBU, a Direct Current (DC)-DC power supply, an active cooling system, and an Alternating Current (AC)-DC (main) supply [36]. The power consumption of a BS is usually modelled by a certain amount of losses and can be influenced by the type of antenna used. The losses includes the feeder, antenna band-pass filters, multiplexers, and matching components. A macro BS has a feeder loss of 3dB [37] which could be mitigated by mounting the PA at the same physical location as the transmit antenna. The feeder loss for smaller BS types is typically negligible.

The digital baseband signal processing is performed by the BBU, which includes digital up/down conversion, Fast Fourier Transform (FFT)/inverse-FFT for OFDM, modulation/demodulation, digital pre-distortion (DPD) (only in downlink (DL) and for large BS), signal detection (synchronisation, channel estimation and equalisation) and channel coding/decoding [36]. As the BS size increases, the digital baseband processing complexity and its power consumption increases with the BS size [38]. The RF functions involve clock/carrier generation and distribution, modulator, mixers, low-noise amplifier, variable-gain amplifier and analog/digital converters. It should be noted that one RF transceiver is installed for each antenna chain, so RF power consumption scales with the number of antenna chains. A relevant portion of the power consumption in a BS is due to the cooling system, DC-DC power

supply and main supply losses which is computed as a fixed power component linearly depending on the total power consumption of other components. It should also be noted that there are no active cooling systems in small cell BSs and cooling is done by natural air. According to [36], it is the BBU that dominates the overall power consumption in small cell BSs (micro/pico/femto) and in the macro BSs it is the PA that dominates the total power consumption due to the high AI losses in macro BSs.

2.4.1.2 Traditional All-in-One BS Architecture

In traditional architectures, BSs have an all-in-one architecture where the radio and baseband processing functionalities are integrated and housed in single cabinet as large as a refrigerator [39]. The BS cabinet is placed in a dedicated room along with all necessary supporting facilities such as power, backup battery, air conditioning, environment surveillance, and backhaul transmission equipment. As shown in Figure 2-12, the antenna module is generally located in proximity (a few meters) to the BBU and connected by coaxial cable which exhibit very high power losses. The all-in-one architecture was popular in 1G and 2G mobile networks and mostly found in macro cell deployments.

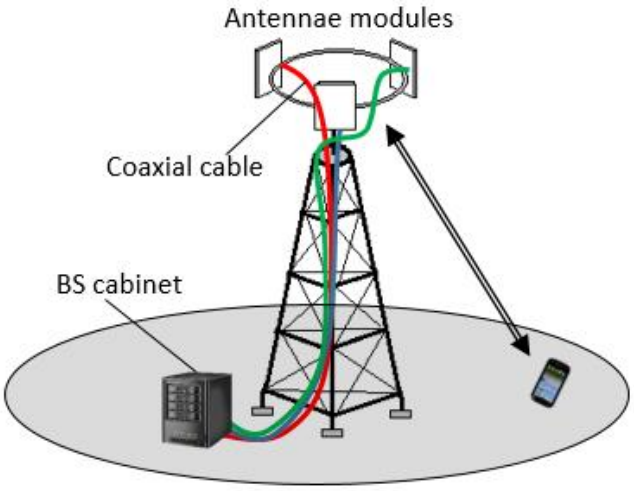

Figure 2-12. Traditional all-in-one BS.

2.4.1.3 Distributed BSs with RRH

In a BS with the RRH architecture, the BS is separated into a radio unit and a signal processing unit, as shown in Figure 2-13. The radio unit called the RRH, performs digital-to-analog conversion, analog-to-digital conversion, power amplification and filtering [40] and is connected to the BBU with an optical fibre cable which prevent losses. The RRH is less intelligent with lower cost, low failure rate, lower downtime and does not require active cooling.

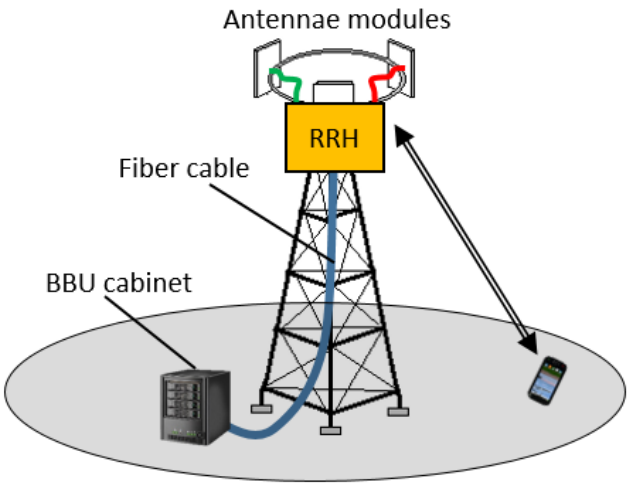

Figure 2-13. Distributed BS with RRH.

The RRHs can be placed up on poles or rooftops, leveraging efficient cooling and saving on air-conditioning in BBU housing. RRHs can be statically or dynamically assigned to BBUs contrary to the traditional RAN. The BBU is used for processing baseband signals from the antennae modules which involves turbo encoding, modulation, pre-coding, FFT, etc. This architecture was introduced primarily for 3G networks and is the most widely used type at this moment.

2.4.2 C-RAN Architecture

In C-RAN, the BS architecture is divided into three parts, the RRH, the fronthaul and the BS cloud/BBU pool as shown Figure 2-14. The BBUs are consolidated at the BS cloud where the baseband processing is performed by virtualised GPPs.

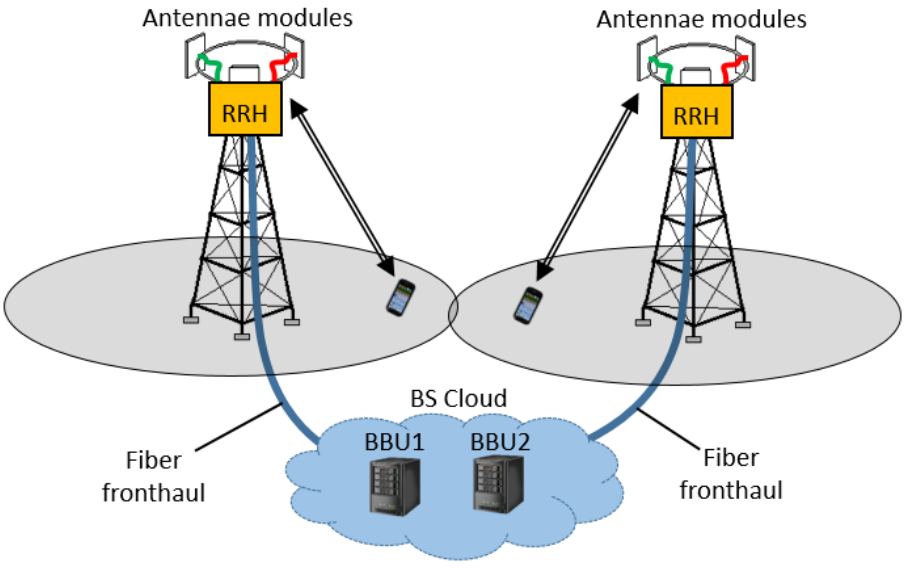

Figure 2-14. C-RAN architecture.

According to the different function splitting between the BBU and RRH, there are two kinds of C-RAN solutions [11]: (i) full centralisation and (ii) partial

centralisation. This thesis will look at the effect that the splitting has on energy efficiency in the network.

2.4.2.1 Full Centralisation

In full centralisation architecture, all baseband (i.e. layer 1 (L1), layer 2 (L2), layer 3 (L3) and operation and management (O&M)) BS functions are located in BS cloud as shown in Figure 2-15.

2.4.2.2 Partial Centralisation

In partial centralisation C-RAN architecture, the RRH integrates not only the radio function but also some of the baseband functions, while all other higher layer functions are still located in BS cloud as shown in Figure 2-16.

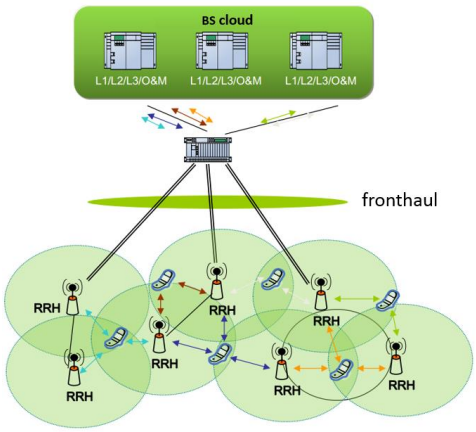

Figure 2-15. C-RAN architecture: full centralisation [11].

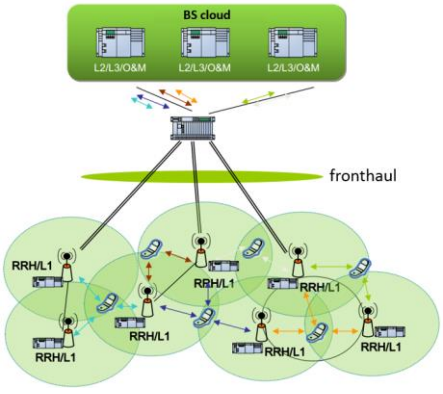

Figure 2-16: C-RAN architecture: partial centralisation [11].

2.4.3 Fronthaul Technologies

Fronthaul is the link between the BS cloud and the RRH and there are various wired and wireless technologies that may be used for this. Digital baseband signal are carried over the fronthaul link, usually using Open Base Station Architecture Initiative (OBSAI) or Common Public Radio Interface (CPRI) standard which has high bandwidth and low latency requirements [11]. Different fronthaul technologies are described as shown in Figure 2-17 . They are divided into wired and wireless technologies.

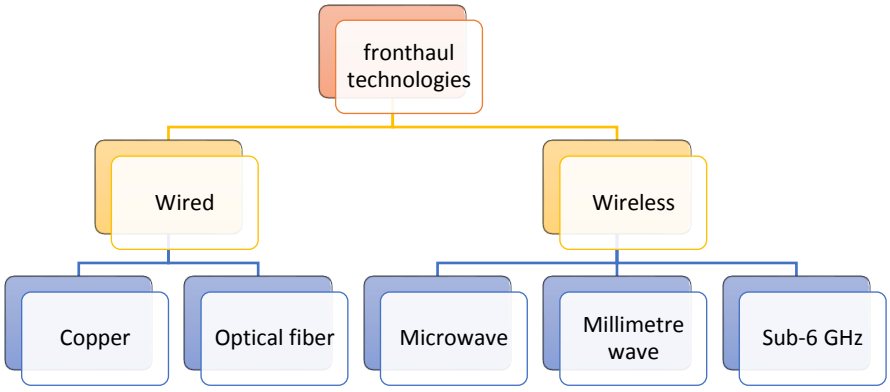

Figure 2-17. Different fronthaul technologies [41].

1. *Wired Fronthaul*

   i. Copper

Considering copper-based solutions, leased T1/E1 copper lines are widely utilised as they can provide suitable support for voice traffic, with deterministic QoS, low latency, and jitter. However, copper lines do not scale easily to provide sufficient bandwidth at distances exceeding few hundred meters to support emerging broadband technologies [6]. Even with 8-pair bonding and vectoring technology, the bandwidth is limited to around 140Mbps [42].

ii. Optical fibre

Optical fibre can provide a 100Gbps+ throughput connectivity for tens of kilometres that can be accomplished using gigabit passive and active optical network technologies [43]. An active optical system uses electrically powered switching equipment, such as a router or a switch aggregator, to manage signal distribution and direct signals to specific customers. A passive optical network uses optical splitters to separate and collect optical signals as they move through the network. Optical fibres are usually deployed in urban and sub-urban areas where very high traffic-carrying capacity is required. Although a fibre-based fronthaul offers long-term support with respect to increasing capacity requirements, this comes at a relatively high CAPEX and costly deployment.

2. Wireless Fronthaul

Various wireless fronthaul solutions exist in terms of the type of propagation, the spectrum used and the network topology. In general, the advantage of wireless backhaul is the freedom from cabling, which is expensive to deploy due to the high costs. Wireless solutions need only equipment's at the small cell and the point of presence offering reduced costs and speed of deployment. The main categories are the following;

i. Microwave

Microwave radio can be viewed as an alternative choice of fronthaul connectivity particularly in areas where a wired connection is not accessible. Microwave transmission works mainly in licensed spectrum (6GHz to 38GHz) and requires Line of Sight (LoS) [44]. Microwave radio can provide capacity of some hundred Mbps.

ii. Millimetre Wave (mmW)

The mmW-radio refers to any RF technology operation in the 30-300GHz range, but it is generally used to discuss 60-80GHz, otherwise called E-band [45]. In this context, several GHz-wide bandwidths are available and can provide multiple Gbps even with low order modulation schemes. On top of high-data rates, mmW radio band can offer superb invulnerability to interference, high security, and the frequency reuse. However, mmW radio requires clear LoS propagation and its range is restricted by the oxygen absorption which strongly attenuates signals greater than 60GHz over distances. Therefore, high gain directional antennas are used in order to compensate for the large free space propagation losses.

iii. Sub-6 GHz

This classification of fronthaul can be seen as a non-LoS category and incorporates carrier frequencies below 6 GHz (3.5GHz licensed and 2.4/ 5.8GHz unlicensed) [41]. Sub-6 GHz fronthaul can be easy to plan and deploy in urban regions, subsequently reducing the cost and duration of small cell network roll out. The 3.5GHz band has emerged as a promising candidate for the dedicated use in small cells. On the contrary, the unlicensed spectrum gives a large amount of freely available bandwidth and it is already heavily used by Wi-Fi hotspots, Bluetooth, and other equipment.

Based on the pros and cons off all the fronthaul technologies, fibre optical cable is chosen in this thesis for its high speed, low latency characteristics.

2.4.4 Key advantages of C-RAN

1. Energy Efficient/Green Infrastructure

C-RAN provides a basis for an eco-friendly architecture which is built on the idea that by centralising the baseband processing at the BS cloud, the number of BS sites can be decreased several folds [46]. Therefore, the power consumption of the air conditioning and other site support equipment can also be largely reduced. In addition, the distance from the subscriber to the RRH can be reduced which allows transmission at low power (e.g. 30dBm) hence interference among RRHs can be reduced while a higher density of RRHs can be deployed. Lower transmission power smaller cells can be deployed while the network coverage quality is not affected. The signal transmission energy will be reduced and this is especially helpful for the reduction of power consumption in the RAN and extends the User Equipment (UE) battery stand-by time. Lastly, because the BS cloud is a shared resource among a large number of virtual BS, it means a much higher utilisation rate of processing resources and lower power consumption can be achieved. When a virtual BS is idle in the evening and most of the processing power is not required, they can be selectively switched off without affecting the constant 24/7 service commitment.

2. Cost-saving on CAPEX and OPEX

Since the BBUs are aggregated into few big rooms, it is much easier for centralised management and operation, saving a lot of the O&M cost associated with the large number of BS sites in a traditional RAN network [39]. Secondly, despite the fact that the number of RRHs may not be reduced in a C-RAN architecture, its functionality is simpler as size and power consumption

are both reduced and they can sit on poles with minimum site support and management. The RRH only requires the installation of the auxiliary antenna feeder systems, enabling operators to accelerate the network construction to gain a first mover advantage. Therefore, operators can get large cost saving on site rental and O&M.

3. Cooperative Processing

In the BS cloud, the virtual BSs can cooperate in a large physical BBU pool and the active UE's signalling, traffic data and Channel State Information (CSI) can be easily shared [47]. It is much easier to implement joint processing & scheduling to mitigate Inter Cell Interference (ICI) and improve spectral efficiency. For instance, cooperative multi-point processing (CoMP) technique in LTE-Advanced can easily be implemented under the CRAN infrastructure.

4. Adaptability to Non-uniform Traffic

Due to the load balancing capability in the BS cloud, C-RAN is also suitable for non-uniformly distributed traffic. Despite the fact that the serving RRH changes dynamically according to the movement of UEs, the serving BBU is still in the same BS cloud. As the coverage of a BBU pool is larger than the traditional BS, non-uniformly distributed traffic generated from UEs can be distributed in a virtual BS which sits in the same BBU pool. As such C-RAN solves the problem of the tidal wave (poor BBU utilisation) and improves BBU utilisation.

5. Smart Internet Traffic Offload

The growing internet traffic from smart phones and other portable devices, can be offloaded from the core network of operators through enabling the smart

offloading technology in C-RAN. The advantages are as follows: reduced back-haul traffic and cost; reduced core network traffic and gateway upgrade cost; reduced latency to the users; differentiating service delivery quality for various applications. A better experience to users is supplied by the service overlapping the core network.

2.4.5 Technical Challenges of C-RAN

C-RAN has many advantages as discussed in Section 2.4.4 from cost saving, increased capacity and energy saving over traditional RAN, however, before deployment by mobile operators, there are some technical challenges in C-RAN as shown in Figure 2-18, that must be addressed.

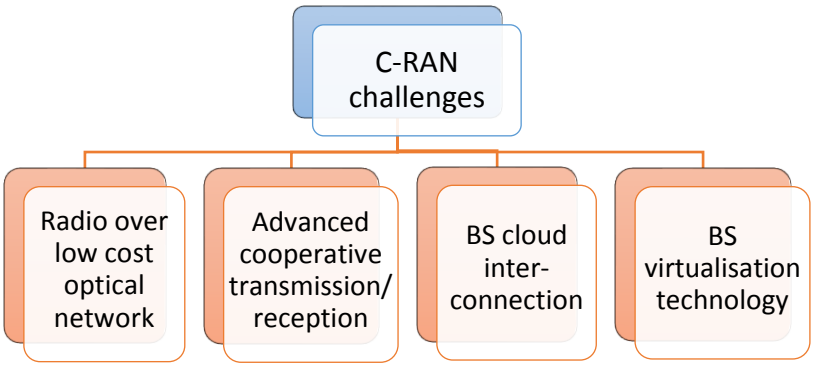

Figure 2-18. Challenges for C-RAN.

2.4.5.1 High Fronthaul Bandwidth and Latency Requirements

Optical fibres between BBU pool and RRHs are required to carry the large amount of tens of Gbps baseband sampling data in real time in the C-RAN architecture [31] from the RRHs to the BS cloud up to 20-40km at very low latencies of 0.1-0.2ms on the fronthaul and with latency jitter of less than 65ns [48].

2.4.5.2 Advanced Cooperative Transmission/Reception

One of the keys to achieve higher system spectral efficiency and mitigate interference in the cellular system is through multipoint joint processing [49] that can make use of special channel state information (CSI) and harness the cooperation among multiple antennas at various cell sites. Joint scheduling of radio resources is also necessary to reduce interference and increase capacity. Both end-user data and uplink/downlink (UL/DL) channel information needs to be shared among virtual BSs in order to support the above cooperative multi-point joint processing algorithms. In order to ensure real time cooperative processing, the interface within the BS cloud interconnecting virtual BSs to carry information should support high bandwidth and low latency. The information exchanged in this interface between virtual BSs includes one or more of the following types: end-user data package, UE channel feedback information, and virtual BSs scheduling information. Therefore, the design of this interface must meet the real time joint processing requirement with low backhaul transportation delay and overheads.

2.4.5.3 Baseband Pool Interconnection

The C-RAN architecture centralises a large number of BBUs within one physical location, thus its reliability is crucial to the whole network. There must be a high bandwidth, low latency, low cost switch network with flexible, extensible topology that interconnects the BBUs in the pool to achieve high reliability in case of BBU failure, in order to recover from error, and to allow flexible resource allocation of BBUS. The digital baseband signals from any RRH can be routed to any BBU in the pool for processing through this switch

network. Thus, any individual BBU failure will not affect the functionality of the system.

2.4.5.4 Base Station Virtualisation Technology

Virtualisation is the partitioning of a physical server into smaller servers to help maximise the server resources [50]. It is important to design virtualisation technologies to consolidate the BBUs into virtual BS entities after the BBUs have been put in a centralised pool. The main challenges of virtualisation are how to create the virtual machines in the BS cloud that will be able to process real time baseband signals and also to perform dynamic baseband processing allocation to deal with the dynamic cell load in system.

2.5 Mobile Edge Computing (MEC)

As smartphones, tablets and wearable devices are gaining enormous popularity, more and more new mobile applications such as face recognition, natural language processing, interactive gaming, and augmented reality are also emerging and attracting greater attention. These kinds of mobile applications are computing intensive and drain a lot of energy in the mobile devices. The mobile devices can hardly cope due to limitations in terms of battery life, storage, processing power and display size [51]. Extending battery life is one of the key requirements by mobile phone users as compared to memory, storage and display size. Hence there is a need for improvement of energy efficiency in mobile devices. One possible approach is to offload the application computation to the remote public clouds such as Amazon Elastic Compute Cloud (EC2) and Windows Azure using Mobile Cloud Computing (MCC) paradigm as shown in Figure 2-19, in which computation is outsourced

to cloud datacentres in the core network. This could help save some amount of energy in the mobile device.

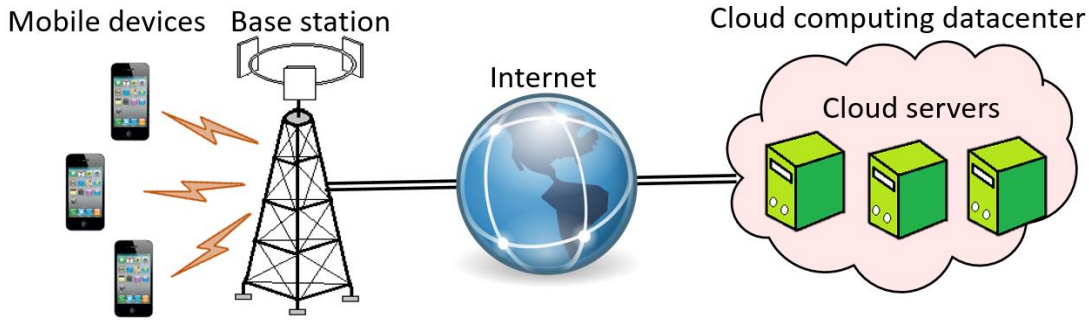

Figure 2-19. An illustration of Mobile Cloud Computing.

These cloud datacentres provide virtually unlimited computation capacity to augment the processors in mobile devices. However, in MCC, the communication between mobile users and remote cloud centres is often over a long distance, adding to the latency in cloud computation. To overcome this limitation, MEC [52], also termed FOG computing [53] was proposed as shown in Figure 2-20.

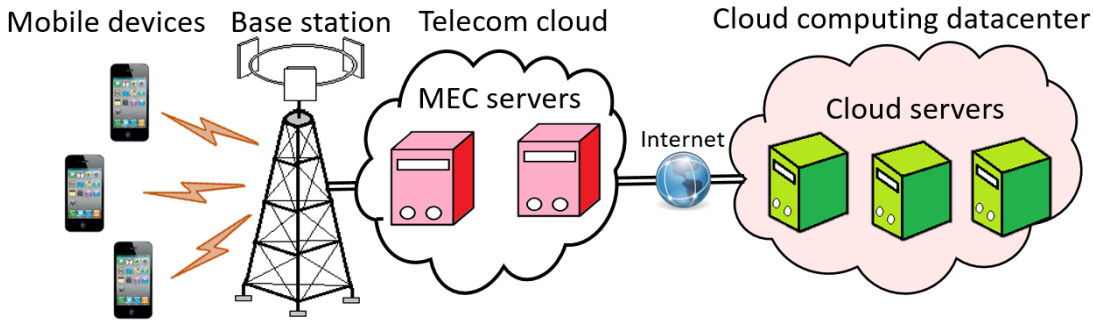

Figure 2-20. An illustration of MEC architecture.

MEC is envisioned as a promising approach to improve the offloading efficiency. In the MEC framework, cloud computing capabilities are provided within the RAN in close proximity to these mobile devices. In other words, with

the aid of MEC, mobile devices are enabled to offload their application tasks to the MEC servers on the edge of the network, rather than utilising the servers in the core network in the cloud datacentres. This MEC paradigm can provide low latency, high bandwidth, high computing agility and improve the energy performance of the mobile devices [52]. Table 2 shows the difference between the MEC and traditional cloud computing.

Table 2. Difference between MEC and traditional cloud computing.

	MEC	**Traditional cloud**
Latency	Low latency	High latency
Resource and service location	At the RAN e.g. BSs, access points, routers and mobile devices.	Dedicated datacentres on the internet.
Mobility	Mobile clients	Mobile and fixed clients
Services	User request and network initiated services	User requested services
Computation capability	Limited computation capability	High computation capability
Service context awareness	Aware of the radio network status and user context	Context available through application reporting

2.5.1 Motivation for MEC

The main motivations for MEC [54] are:

i) Computation offloading

ii) Dynamic content optimisation

iii) Mobile big data analytics

1. Computation offloading

As the recent exposure of wearable devices, low processing power IoT devices, and smartphones becomes prevalent, computing intensive

applications cannot be performed in the device itself due to limited storage and computation and processing. This problem can be solved by splitting the application into small tasks then some of the tasks can be performed in the cloud provided that the delay deadlines are met. The tasks can be offloaded to the edge server in closer proximity to the device itself. The key of challenges in MEC computation offloading are: i) How to split the application and ii) How to identify whether a task should be offloaded or not.

2. Dynamic content optimisation

To fulfil the customer's expectations, traditional content optimisation is performed at the web hosting site where the content optimiser uses the user's web surfing history stored in the database [55]. These traditional methods incur some delays and inaccuracies. In MEC, content optimisation can be done based on the user's context aware information dynamically where the content optimiser can be located at the edge server. The optimiser in MEC can acquire accurate RAN information like network status and network load dynamically and these information can be used for optimisation [54]. This MEC content optimisation can improve network performance, QoE and new services can be added easily.

3. Mobile big data analytics

Big data is a collection of large volumes of both structured and unstructured data and big data analytics is the process of analysing such big data for better decisions and gaining strategic business insights [54]. Data collection from the edge devices in big data analytics are transferred to the core network and this process takes high bandwidth and latency. The MEC platform can be used to perform big data analytic at the edge of the network thereby saving large

amount of bandwidth required. After the analysis, the results can be sent to the core network. As such, the scenario will reduce bandwidth consumption and improve the network latency.

2.5.2 Taxonomy of MEC

The reader is directed to comprehensive surveys on MEC in [54, 56]. Figure 2-21 shows the taxonomy of MEC which is based on: i) Characteristics, ii) Actors, iii) Access technologies, iv) Applications, v) Objectives, vi) Computation platforms and vii) Key enablers. These parameters are described below.

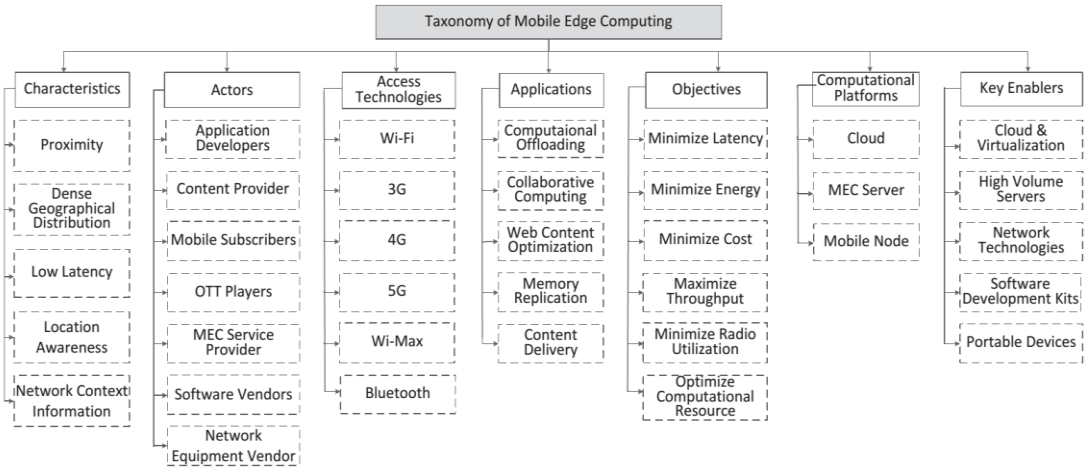

Figure 2-21. The taxonomy for MEC [54].

1. Characteristics

The key characteristics of MEC are as follows:

i) **Proximity:** Where the mobile device is closer to MEC server in the RAN. The MEC server can also be another mobile device through device to device (D2D) communication forming a mobile cloud. Since, edge server is nearby to devices; it can extract device information and analyse user's behaviour to improve services.

ii) **Dense geographic distribution:** Where the MEC hosts the cloud computing services at the edge network which is located at numerous locations. As such, these geographically dispersed infrastructure contribute to the MEC in many ways.

iii) **Low latency:** It takes a long time when transmitting data to the core network but with MEC, transmission is faster as the MEC servers are closer to the mobile devices.

iv) **Location awareness:** Where application developers can use the user's locations to provide context aware services. User mobility patterns can be collected easily at the MEC to predict future network status.

v) **Network context information:** Where real time RAN information such as subscriber location, radio condition, network load etc. is used to provide context related services to the mobile subscriber. RAN information is used by the application developers and content providers thus improving user satisfaction.

2.6 Concluding remarks

This chapter provided an overview of RANs by surveying the five generations of cellular networks (1G, 2G, 3G, 4G and 5G) in Section 2.2. It has been noted that the advancement of RAN technologies is to increase capacity. The challenges of conventional RAN systems were discussed in Section 2.3 including poor resource utilisation, high BS energy consumption, increased cost etc. Furthermore, new types of RAN architecture based on cloud computing technology called C-RAN was discussed in Section 2.4 where the evolution, architecture, motivation and technical challenges were presented. Another cloud computing based RAN architecture called MEC was introduced

in Section 2.5 as a promising way of improving energy consumption in mobile devices. The presented RAN systems like C-RAN and MEC are a road towards 5G.

3 Energy Efficiency in C-RAN and Mobile Edge Computing– A Literature Review

3.1 Introduction

Over the last two decades, there have been considerable efforts to improve the cellular system's capacity by deploying more BSs, multiple antennas, and so on. Nevertheless, the disadvantage is that there is an enormous energy consumption associated with the wireless communications infrastructure. Also, the proliferation of energy hungry mobile applications has resulted in high energy consumption in mobile devices which led to the proposal of MEC. This chapter presents the need of energy efficient cellular networks both in C-RAN and in MEC and their energy expenditures. The goal of this chapter is to present the current techniques used within C-RAN and MEC in increasing energy efficiency.

3.2 Energy Saving within C-RAN

In order to secure both environmental and financial benefits, there is considerable scope for significant improvement in energy efficiency of cellular networks. Improvement of energy efficiency in cellular networks involves energy reduction of all elements, such as mobile core networks, mobile switching centres, BS, backhaul networks and UE. In fact, BSs are the most energy consuming elements of the cellular system, and are often under-utilised during low traffic periods. A number of strategies have been proposed to reduce the carbon footprint of BSs as described in [12], ranging from the use of renewable energy sources, improvement of hardware component,

optimum network deployments, optimisation of resource management and transmission techniques (e.g. multi input multi output (MIMO), coordinated multipoint transmission etc.), cell zooming techniques, to the adoption of BS sleep modes. Given these wide number of research works in energy efficiency of cellular wireless networks, this thesis will discuss the major works relevant to C-RAN. Several international research projects on green communication networks have been initiated under different international research platforms in recent years as follows.

1. *Mobile Virtual Centre of Excellence's Green Radio (MVCE)*

   The MVCE project [57] was established in 2009 to show how significant energy savings may be obtained in future wireless systems. The vision for the project is to specify an LTE compliant BS that is able to operate at much lower overall consumption, possibly sufficiently low to enable operation from renewable sources locally generated (e.g., solar or wind). The project aims to reduce energy consumption in BSs through (i) improved resource allocation strategies, (ii) proposing interference management and mitigation (iii) and energy efficient routing and multi-hop through the use of relays to exchange information between the BSs and mobile terminals. However, the MVCE project only considers stand-alone BSs and does not consider the C-RAN architecture which is considered in this thesis.

2. Energy Aware Radio and Networks Technologies (EARTH) Project

The EARTH project [58] is a major European research project which was initiated in 2010 with 15 partners from 10 countries with the main objective of

reducing the energy consumption within the mobile network by 50%. EARTH claims to reduce energy consumption through the following means:

- Deployment: Mixture of cell sizes, use of repeaters, multi RAT deployment.
- Radio resource management (RRM) algorithms: cooperative scheduling, interference coordination, joint power allocation, coverage adjustment, multi-RAT coordination.
- Disruptive approaches: multi-hop transmission, terminal to terminal transmission.

However,similar to the MVCE project, the EARTH project does not consider C-RAN but target energy reduction in stand-alone BSs.

3. Green Touch Project

Green Touch [59] was initiated in 2010 with the main goal of delivering architecture, specifications and technologies needed to reduce network energy consumption by 90% compared to 2010 levels by adopting the following:

- Separating signalling (continuous and full coverage service) and data network (on demand service);
- Flexible power model of future BSs based on operation modes, component configuration, hardware technology and BS architecture;
- Fundamental trade-off between energy efficiency and spectrum efficiency, and between service delay and energy consumption.
- Use of large scale antenna systems.

However, Green Touch also does not consider the cloud based BSs architectures for 5G.

4. Towards Green 5G Mobile Networks (5GrEEn) Project

5GrEEn was launched in 2013 with the main objective of designing a green 5G mobile networks, a clean slate solution for environmentally friendly mobile networks of the future by:

- Integrating mobile access with fibre
- Finding trade-offs between energy efficiency and QoE.
- Optimising the trade-off between QoE and energy-efficiency in converged mobile access networks
- Traffic adaptive solutions and deployments strategies in 5G

Nevertheless, 5GrEEn does not consider BS sharing which can significantly improve the BS utilisation.

5. Green Net Project

The Green Net project [60] is focused on the analysis, design and optimisation of energy efficient wireless communication systems and networks. The energy efficiency is done by adopting the following:

- Novel energy efficient and low complexity Physical layer technique: Spatial modulation for generalised MIMO
- Energy efficient scheduling, sleep modes and interference management for heterogeneous networks
- Data aggregation schemes
- Network/channel coding

- Energy efficient cooperative spectrum sensing for cognitive radio

The Green Net Project does also not consider the sharing of BS and the switching off underutilised baseband BBUs.

Major mobile operators have already joined hands to initiate several experimental projects to explore C-RAN and energy efficiency in cellular networks. However none of these projects consider dynamic sharing of BBU servers in the cloud and dimensioning the BBUs with traffic load. Table 3 highlights a broad overview of the major industrial initiatives pertaining to C-RAN and energy efficiency. The NGMN alliance's P-CRAN [61] project, European Union (EU)'s Mobile Cloud Networking (MCN) [62], FP7-based High Capacity Network Architecture with RRHs and Parasitic antenna arrays (HARP) [63], iJOINT [64] and CROWD [65] projects are major initiatives for C-RAN design and implementations. Prototyping and field trials of C-RAN are already underway by major cellular operators and vendors, like China Mobile [11], Ericsson [66], NEC [67] and Korea Telecom (KT) [68]. Cellular vendors and operators, like ZTE [69], China Mobile and KT estimate around 70% power savings by the deployment of C-RAN.

Table 3. Major industrial initiatives in C-RAN.

Major Projects	**Field Trials**	**Energy Efficiency**
PCRAN [61],MCN [62], FP-7/HARP [63], iJOINT [64], CROWD [65].	China Mobile [11], Ericsson [66], NEC [67], KT [68].	ZTE [69], China Mobile [11], KT [68].

In the following subsections, the different energy saving techniques within C-RAN will be described as shown in Figure 3-1.

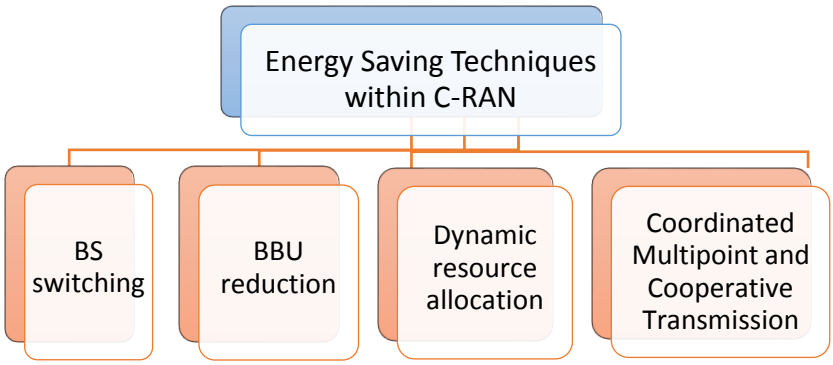

Figure 3-1. C-RAN energy saving techniques.

3.2.1 BS Sleeping

Enhancing the energy efficiency of cellular networks largely depends on the effective operation and management of the BS. Turning off some BSs during low-traffic periods, called BS sleeping, is being considered as one of the most promising techniques. This can be confirmed from the emergence of a large number of BS switching schemes in recent years which has drawn more and more attention in literature [70-75].

Depending on the instantaneous traffic conditions within cells and their spatial positions in a network, the number of active BSs (and hence the overall energy consumption) is dynamically minimised, while the remaining active BSs provide the radio coverage and service provisioning (required QoS) for all users within the network [12]. In other words, the objective is to keep the minimum number of active BSs within the network at any instant of time such that the network energy consumption is low while the QoS is met. This is also the principal research focus of this thesis which however looks at further

reductions in energy usage by the use of virtualised BSs adopting a C-RAN architecture.

The authors in [4] investigates the approaches to improve energy efficiency of a centralised super BS implementing BS switch off mechanism in the radio side and sharing of housing and BBU's on the cloud side. However, the authors did not consider energy consumption brought about by the fronthaul in C-RAN. The authors in [76] proposed a BS switch off algorithm called fixed progressive dynamic switching off (*F-PDSO)*. In *F-PDSO*, only the pico BSs are switched off while the macro BS maintains coverage. In [77], the authors proposed an analytical energy model of a computational-resource-aware virtual BS in a cloud-based cellular network architecture. The authors consider the energy-delay trade-offs of a virtual BS considering BS sleeping mode in GPP platforms. The authors in [78] developed a game theoretic approach and a testbed for switching off RRHs in C-RAN during off-peak hours and simulation results shows that their scheme improves the average energy efficiency by 35% while offering 30% daily energy saving.

3.2.2 BBU Reduction

One of the ways of reducing energy consumption in C-RAN is to reduce the number of BBUs in the cloud by consolidating the workload from RRHs into fewer number of BBUs during low traffic periods. Authors in [33] proposed a BBU reduction scheme for C-RAN that allocates BBUs to RRHs based on the imbalance of subscribers in office/residential areas. Even though the reduction of the number of BBUs required is achieved, the model depend on an inefficient random allocation of RRH traffic to BBUs. Random allocation of RRHs to BBUs means that the model does not first check the utilisation of

BBUs as to whether the BBU is highly loaded or not but rather allocates the RRH to any BBU and this can cause overloading of BBUs. S. Namba *et al.* in [79] proposed a BBU reduction network architecture called Colony-RAN due to its ability to flexibly change cell layout by changing the connections of BBUs and RRHs in respect to traffic demand. Because traffic demand fluctuates dynamically for each time and location, Colony-RAN can drastically reduce the number of BBUs due to a statistical multiplexing effect by establishing one BBU that can connect to several RRHs until the baseband resources are exhausted. However, this proposed method depends on randomly allocating RRHs to BBUs.

The authors in [80] proposed a BBU virtualisation scheme that minimises the power consumption of the BBU pool. The BBU virtualisation problem is formulated as a bin packing problem, where each BBU is treated as a bin with finite computing resources, expressed in million Operations per Time Slot (MOPTS). The dynamics of the cell traffic load is treated as an item that needs to be packed into the bins with the size equal to the computing resources, required to support the traffic load. However, the proposed scheme does not consider or quantify the workload representing the items from the RRHs to be packed into bins. Also the authors do not specify the type of resource constraints considered in the BBUs such as CPU, random access memory (RAM), etc.

On the other hand, the authors in [81] proposed a cloud provisioning model for an LTE network consisting of one macrocell and variable numbers of femtocells in the form of RRHs and their model allowed for a smooth coordination of densely heterogeneous networks through central core

management. The authors proposed a graph colouring model that switches off the virtual BBUs clusters with no or low traffic, enabling significant power savings that are obtained from both fronthaul and backhaul sides. The authors in [82] proposed a RAN as a Service (RANaaS) framework for saving energy in C-RAN by sharing BBUs in the BS cloud and switching off idle BBUs. However, the authors assume all the BBUs have the same CPU utilisation irrespective of traffic load processed in that BBU.

3.2.3 Coordinated Multipoint and Cooperative Transmission

3GPP proposed to use CoMP to facilitate cooperative communications across multiple transmission and reception points (e.g., cells) for the LTE-A system [83]. In CoMP operation, multiple points coordinate with each other in such a way that the transmission signals from/to other points do not incur serious interference or even can be exploited as a meaningful signal. Liming Cheng *et al.* in [47] focused on the spectral efficiency and energy efficiency of C-RAN implementation which was achieved by cooperative transmission among RRHs whereas energy efficiency performance was improved through proposed computationaly efficient pre-coding scheme. The authors in [84] investigate the cooperative transmission design for C-RAN considering fronthaul capacity and cloud processing constraints. The authors consider the joint transmission scheme where the baseband signals and precoding vectors are processed and calculated by the cloud. The authors in [85] address the energy efficiency issue in downlink C-RAN using joint RRH selection and user association scheme with the goal of network energy minimisation while satisfying the required data rate by deactivating the RRHs if there are no associated users. The authors in [86] have presented an efficient and low-

complexity algorithm for downlink coordinated transmission in C-RAN where the aim is to minimise the total transmission power subject to constraints on transmission powers, fronthaul capacity and required QoS of users. Numerical results have illustrated the efficiency of their proposed algorithms and the impacts of different parameters on the network performance. The authors in [87] proposed HetNet in C-RAN as a promising new paradigm to achieve high spectral efficiency and the energy efficiency performance through the combination of cloud computing and HetNets. The authors have surveyed key large-scale cooperative processing and networking techniques, including cloud-computing based CoMP and cooperative radio resource management.

3.2.4 Dynamic Resource Allocation

The authors in [88] reduced the network cost and energy consumption in C-RAN by dynamically allocating centralised BBU resources to RRHs depending on the traffic conditions, however, the power consumption on RRHs and BBUs are assumed to be static and are independent of traffic load which is not realistic. Authors in [89] proposed a C-RAN system using virtualisation technology on GPPs where BBUs are dynamically provisioned according to traffic load. In [67], L. Cheng *et al.* developed an energy efficiency C-RAN system with a reconfigurable backhaul that allows four BBUs to connect flexibly with four RRHs using radio-over-fibre technology. The backhaul architecture allows the mapping between BBUs and RRHs to be flexible and changed dynamically to reduce energy consumption in the BBU pool. However, the paper assumes static user traffic whereas in reality BS traffic is dynamic. The authors in [90] proposed a dynamic resource allocation in C-RAN with real-time BBU-RRH assignment considering the constraint in

transmission power and Signal to Interference Plus Noise Ratio (SINR) for the UE.

In all the proposed schemes above, there are common deficits that are noticed as follows:

- BBUs are still decentralised to their respective coverage areas and remain active even when idle as such, the authors consider standalone BBUs that are never shared with each other.
- There is no switching off of BBUs in their proposed schemes since the authirs do not consider further minimising energy by reducing the number of BBUs in the BBU pool.
- The authors did not consider minimising energy consumption at the radio side and in the fronthaul.
- The authors do not consider the energy savings of the virtual BS model as it scales with traffic load.

As such, our work will address all this issues for an efficient 5G system.

3.3 Energy Saving within MEC

The typical applications of MEC includes mobile commerce, mobile learning, mobile healthcare, mobile gaming and other practical applications like social networking, showing maps, storing images and video [91]. These applications when processed in the mobile devices consume a lot of battery power, as such energy efficient schemes for saving energy in mobile devices by offloading application to the MEC cloud are presented in this section.

The authors in [92] investigate a green MEC system and developed an effective computation offloading strategy. The execution cost, which addresses both the execution latency and task failure, is adopted as the performance metric. A low-complexity online algorithm is proposed, namely, the Lyapunov optimisation-based dynamic computation offloading algorithm, which jointly makes the offloading decision, the CPU-cycle frequencies for mobile execution, and the transmit power for computation offloading. A unique advantage of the Lyapunov is that the decisions depend only on the current system state without requiring distribution information of the computation task request and of the wireless channel. Nevertheless, the authors assume that the battery capacity is sufficiently large which is impractical, also the authors ignore the execution delay caused by the MEC server.

Chen L. *et al.* in [93] addresses the challenge of incorporating MEC into dense cellular networks, and propose an efficient online algorithm, called ENGINE (ENergy constrained offloadINg and slEeping) which makes joint computation offloading in order to maximise the QoS while keeping the energy consumption low. However, the authors assume that traffic among BSs is equally distributed whereas traffic is randomly distributed in reality. Zhang K. *et al.* in [52] proposed an energy efficient computation offloading (EECO) mechanisms for MEC in 5G HetNets. In EECO, an optimisation problem was formulated to minimise the energy consumption of the offloading system, where the energy cost of both task computing and file transmission are taken into consideration. However, the authors do not show the impact on the response time of an application offloaded to the MEC server.

The authors in [85] study the multi-user computation offloading problem for MEC computing in a multi-channel wireless interference environment by formulating the distributed computation offloading decision making problem among mobile device users as a multi-user computation offloading game. Numerical results corroborate that the proposed algorithm can achieve superior computation offloading performance and scale well as the user size increases. However, the application to be offloaded is assumed to be atomic (the application cannot be divided into tasks), hence the whole application code is sent to the MEC server which can incur more transmission costs.

Deng M. *et al.* in [94] investigated the computation offloading decision making problem in a multi-cell MEC scenario. The authors proposed an adaptive sequential offloading game approach and designed a multi-user computation offloading algorithm where the mobile users sequentially make offloading decisions based on the current interference environment and available computation resources. Numerical results show that their proposed algorithm can achieve efficient performance and scale well as the network system size increases. Nevertheless, there is also a lack of a global control manager to manage the requests from the mobile users.

Beck et al. in [95] proposed MEC enabled Voice over LTE (ME-VoLTE) architecture to reduce battery consumption of mobile devices during video call and provide a communication protocol for negotiating the offloading strategy. Here, the process of video encoding during video call processing is offloaded at the MEC edge server. Nevertheless, the authors consider a MEC server at the edge of a BS which has limited processing capacity.

The downfalls of the above MEC energy saving schemes in the mobile devices is that the MEC server at the edge of the network (at the BS) are of limited capacity and storage, hence too many requests from users can overload the MEC server in next generation networks where there will be many applications for mobile users. There is also a lack of a global control manager to manage the requests from the mobile users which leads to poor performance.

3.4 Concluding Remarks

This chapter has looked at the various energy saving schemes in C-RAN which fall primarily under the categories of BS switching, BBU reduction, CoMP and dynamic resource allocation. The key drawbacks with the existing techniques is that they only consider static network traffic, do not consider power consumption in the fronthaul and only consider standalone BBUs which are not shared to further reduce energy consumption. The chapter also presented the energy saving techniques for MEC proposed in literature which mainly look at offloading computation hungry application tasks to the MEC server while meeting delay deadlines of the application. The downfalls of these conventional MEC energy saving schemes in the mobile devices are that the MEC server at the edge of the network (at the BS) are of limited capacity and storage as such too many requests from users can overload the MEC server in next generation networks where there will be many applications for mobile users. Hence there is need for new schemes that will further reduce energy consumption within the C-RAN at the radio side and in the BBU pool and also within the mobile device.

4 Proposed Energy Efficient 5G C-RAN Framework

4.1 Introduction

This chapter introduces the proposed energy efficient 5G C-RAN framework termed H-C-RAN as it combines HetNets and C-RAN. In the proposed H-C-RAN framework, energy saving is done in three ways as shown in Figure 4-1.

i) **At the radio side of C-RAN:** A novel BS sleeping mechanism is implemented that incorporates CAC;

ii) **At the cloud side:** Baseband processing workload consolidation via virtual BBU placement is employed along with an advanced CAC scheme implemented in the BS cloud to improve QoS; and

iii) **At the mobile device side**: MEC paradigm is implemented in H-C-RAN framework to save energy in the mobile device by computation offloading where processing and energy hungry applications are split into tasks which are then executed in the MEC server in the BS cloud.

These key blocks of the proposed framework are discussed in the following sub sections in details.

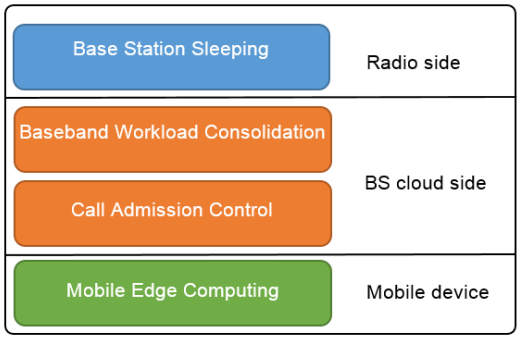

Figure 4-1. Proposed framework block diagram.

4.2 System Model and Problem Formulation

4.2.1 H-C-RAN Framework Architecture

The architecture of the proposed H-C-RAN framework as shown in Figure 4-2, extends the LTE-A architecture. At the cloud side, full C-RAN centralisation as described earlier in Section 2.4.2 is implemented, where 100% of the baseband processing is centralised in the BS cloud including physical layer one(PHY-L1), MAC, radio link control (RLC), packet data convergence protocol (PDCP) and other higher layers like radio resource control (RRC). The mobility management entity (MME) and serving gateways (S-GW) can also be located in the cloud but this is not the target for the thesis which only concentrates on the RAN aspects of LTE-A involving baseband processing.

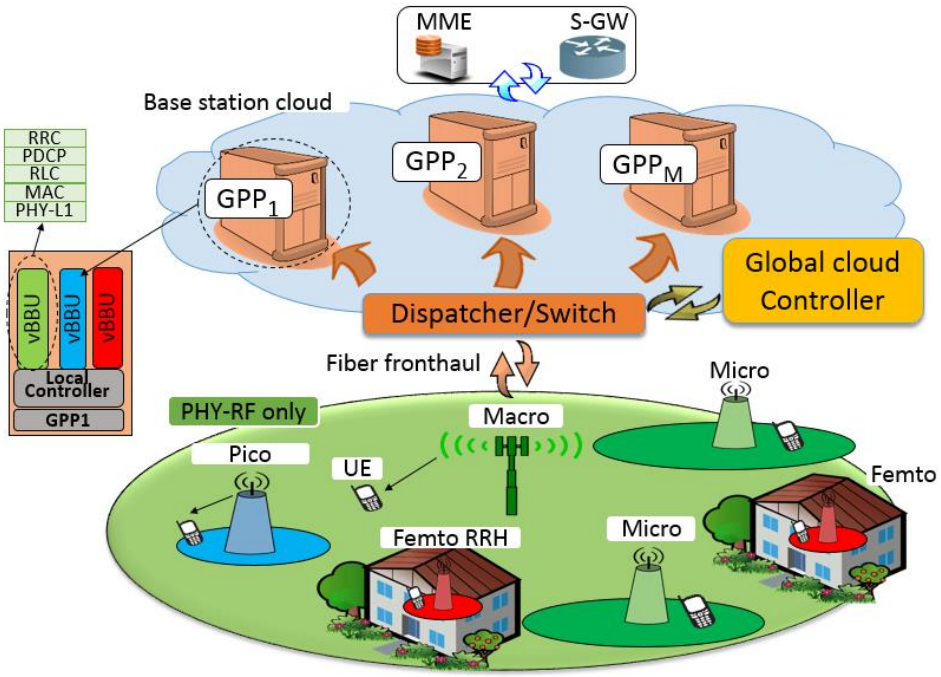

Figure 4-2. H-C-RAN framework architecture.

The baseband is processed on virtualised GPPs which are affordable and can process any signal from any RRH. The GPPs are virtualised into VMs used for baseband processing herein termed vBBUs. The BS cloud is linked to the

radio side by high bandwidth, low latency optical fibre cables. A dispatcher or a switch is used to distribute the baseband signals from the RRHs to the GPPs. The global cloud controller (GCC) is the main module of the H-C-RAN framework located in the BS cloud and it is where the BS sleeping/switch off, baseband workload consolidation and CAC are located whereas the MEC framework is located in the mobile device.

4.2.2 System Model

Consider an LTE-A downlink transmission based on orthogonal frequency division multiple access (OFDMA) of a two-tier H-C-RAN consisting of macro cells overlaid by pico cells. In the proposed H-C-RAN, the macro BSs and pico BSs are replaced by less intelligent RRH as in [96]. The RRH of macro cell (MRRH) and the RRH of pico small cell (PRRH), have the same coverage and height as the macro and pico BS respectively. Also the separation of control and data signalling [22] have been adopted in this thesis where the MRRH mainly provide control signalling (network access provision) whereas the PRRH provide data transmission as such signalling overheads are reduced in the PRRHs. The H-C-RAN consists of a set of RRHs $\boldsymbol{R} = [RRH_j : j = 1\ldots, N_R]$ and a set of users equipment's (UE) $\boldsymbol{U} = [k : k = 1\ldots, N_{UE}]$ where N_R and N_{UE} denote the total number of RRHs and the total number of active users in the network respectively. Each RRH covers a cell and is serving a set of users $\boldsymbol{U}_j$. Denote $N_a : a = [1,2]$ such that N_1 denotes the total number of MRRHs and N_2 denotes the total number of PRRHs such that $N_1 + N_2 = N_R$. $\boldsymbol{R}_p = [PRRH_x : x = 1, \ldots, N_2]$ denotes the set of PRRH. Each UE k will associate with a serving RRH. Cell association is decided by criterion of cell

range expansion (CRE) to be explained later in Section 4.2.3. The system bandwidth, B, is divided into N_{ch} which can be the number of physical resource blocks (PRBs) each with the bandwidth w. PRBs are the basic downlink resource allocation units for data transmission and any PRB can be allocated to only one UE within a cell. Link adaptation is considered and channel quality indicator (CQI) from each user is used to determine a specific modulation and coding scheme, which denotes the amount of bits per symbol. Co-channel deployment is considered; hence, all channels are simultaneously allocated by both MRRHs and PRRHs to their served UEs. It is assumed that mobile users in the same cell transmit over orthogonal channels, whereas users of different cells may interfere against each other.

In the BS cloud side, a full centralisation C-RAN system is considered where virtualised GPPs are used as baseband processors. The VMs running on the GPPs are denoted as vBBUs and each RRH has its own specific vBBU. Some of the VMs are used for processing the MEC applications tasks which will be explained later in Section 4.7. Define a set of GPPs in the BS cloud as $\boldsymbol{M} = [GPP_i : i = 1, 2, \ldots, N_G]$ where N_G is the total number of active GPPs within the cloud to be minimised. Therefore assume a set of vBBUs as $\boldsymbol{V} = [vBBU_j : j = 1, 2, \ldots, N_v]$ where N_v is the total number of vBBUs in the BS cloud.

4.2.3 User Association

Typically in LTE-A cellular networks, when a UE device has to select a suitable cell for association, it chooses the one with maximum pilot power, i.e. maximum reference signal received power (RSRP). However, this criterion cannot be used in HetNets because in the expanded region of PRRHs in

HetNets, the power imbalance between DL and UL leads to a mismatch between the UL and DL handover boundaries. Therefore, associating a UE to the RRH which provides the strongest DL RSRP may not always be the best strategy and is not an efficient way of network resource usage. CRE is an alternative cell association that has been widely discussed in literature [97, 98] and will be adopted in this thesis. In CRE, generally, a positive bias is added to the DL RSRPs of MRRHs pilot signal to increase PRRH's DL coverage footprints, thus compensating for the DL/UL mismatch. As such, the UE become associated to the preferred RRH, RRH_{pref}, such that;

$$\mathrm{RRH}_{\mathrm{pref}} = \arg\max(\mathrm{MacroRSRP}, \mathrm{PicoRSRPO} + \mathrm{bias}) \quad (1)$$

4.2.4 Resource allocation

In LTE, the smallest radio resource unit that the scheduler can assign to a user is a PRB [99]. The PRB contains 12 adjacent OFDM subcarriers with an inter-subcarrier spacing of 15 kHz. Each PRB has a time duration of 1 ms, corresponding to 12 or 14 orthogonal OFDM symbols depending on whether an extended or normal cyclic prefix is utilised. The assignment of PRBs to users is done by the scheduler, which takes decisions for each subframe, i.e., every 1 ms. It is assumed that the subcarriers constituting a single PRB have the same fading and hence the channel gain on the subcarriers of a single PRB is considered to be the same. Moreover, the fading is assumed to be independent identically distributed (IID) across PRBs. A baseline CAC algorithm is implemented where the MRRH can only accept up to 100 users and the PRRH can only accept up to 8 users simultaneously due to its smaller

coverage area. Users more than these values are dropped. Apart from that, communications between the RRH and UE can be disrupted due to two reasons [12]:

a. Connection blocking: due to the lack of required radio resources for newly arrived users;
b. Signal outage: due to the received signal strength i.e., the received SINR being less than a predefined threshold.

The probabilities for both these reasons are important parameters that reflect the performance of the framework. Connection blocking probability measures performance whereas signal outage probability measure the link quality between the transmitter and receiver. When a cellular system is designed, the blocking probability should be less than a particular predetermined threshold value, $P_{th} = 5\%$ [12] to ensure a minimum QoS.

4.2.5 Achievable Downlink Datarate for Users

In LTE-A, the different PRBs allocated to a UE can have different modulation and coding schemes (MCS). In this thesis different MCS schemes are considered such as 4QAM, 16QAM and 64QAM. The UE measures the SINR which is converted to CQI value that ranges from 0 to 15 and the CQI is then reported to the BS cloud. The UE sends CQI feedback to the BS as an indication of the datarate which can be supported by the downlink channel. The UE determines CQI such that it corresponds to the highest MCS allowing the UE to decode the transport block with error rate probability not exceeding 10%, that is, the ratio of the number of bits that are received in error within the transport block to the total number of bits should not exceed 10%. In this

thesis, automatic repeat request (ARQ) was not taken into consideration. However, the minimum datarate for the UE has to be satisfied for the user to be served by a particular RRH or else there is an outage. Assume that a UE k is served by RRH_j, and $\boldsymbol{I}^{j}_{PRB,k}$ is the set of PRBs allocated from RRH_j to UE k, then the achievable rate of k can be expressed as:

$$r^{k}_{UE} = \sum_{n \in \boldsymbol{I}^{j}_{PRB,k}} w.\log_2(1 + SINR^{j}_{k,n}) \tag{2}$$

$$SINR^{j}_{k,n} = \frac{P^{j}_{k,n} G^{j}_{k,n}}{I^{j}_{k,n} + \sigma^{2}_{k,n}} \tag{3}$$

$$I^{j}_{k,n} = \sum_{j'=1, j' \neq j}^{N_{UE}} \left(\sum_{k' \in \boldsymbol{U}_{j'}} \alpha^{j'}_{k',n} \right) * P^{j}_{k,n} G^{j'}_{k,n} \tag{4}$$

where $SINR^{j}_{k,n}$, $G^{j}_{k,n}$ and $P^{j}_{k,n}$ denotes the SINR on the n^{th} PRB of UE k, the average channel gain from RRH_j to UE k on the n^{th} PRB and power consumption of the PRB. The channel gain involves path loss and shadowing. Only large-scale fading of the channel model is considered, and small-scale fading is omitted since the main focus of this thesis is not on channel modelling but large scale fading for RRM like BBU reduction and BS sleeping. The power of additive white Gaussian noise in the n^{th} PRB is denoted as σ^2. The variable $I^{j}_{k,n}$ in (4) denotes the interference on the n^{th} PRB that a UE receives. The scheduling indicator $\alpha^{j'}_{k',n} \in [0, 1]$ denotes the association of the n^{th} PRB of user k in $RRH_{j'}$. RRH_j, will cause interference to UE k if the RRH is occupied by a UE otherwise there is no interference.

4.3 Energy Consumption Model

4.3.1 EARTH Model

The EARTH project [82] defined a power model for distributed LTE-A system that investigated how the power consumption of distinct components of several evolved NodeB (eNodeBs), such as PA, baseband engine, RF transceiver, (DC)-DC converter, main supply, and active cooling, depends on the transmission bandwidth, the transmission power, and the number of radio antennas. Two power consumption models were proposed in [82]; a linear model and a component based model. For the linear model, the power consumption of an eNodeB, P_{eNodeB}, can be approximated as [100]:

$$P_{eNodeB} = \begin{cases} N_{TRX}(P_0 + \Delta_p.P_{max}.\rho_j); & \text{if } 0 < \rho_j \leq 1 \\ P_{sleep}; & \text{if } \rho_j = 0 \end{cases} \quad (5)$$

where N_{TRX} and P_0 denote the number of transmitter antennae and the static power consumption of the eNodeB respectively. The term Δ_P denotes a power gradient variable of a particular BS type, P_{max} denotes the maximum transmission power when cell load is 100%. The scaling parameter ρ_j is the normalised cell traffic load of the j^{th} eNodeB, where $\rho_j = 1$ indicates a fully loaded system, e.g. transmitting at full power and full bandwidth, and $\rho_j = 0$ indicates an idle system. P_{sleep} is power consumption when the eNodeB is in sleep mode with 0% traffic load. Thus, the total power consumption of the entire network for the baseline distributed LTE-A system $P_{baseline}^{linear}$ is then formulated as:

$$P_{baseline}^{linear} = N_{TRX} . \sum_{j=1}^{N_R} (P_0 + \Delta_p . P_{max} . \rho_j); \text{ if } 0 < \rho_j \leq 1 \quad (6)$$

On the other hand, the general BS component based power consumption model $P_{baseline}^{comp}$ for an eNodeB can be formulated as [100]:

$$P_{baseline}^{comp} = N_{TRX} . \frac{\frac{P_{out}}{\eta_{PA}(1-\sigma_{feeder})} + P_{RF} + P_{BB}}{(1-\sigma_{DC})(1-\sigma_{MS})(1-\sigma_{cool})} \quad (7)$$

where P_{RF} and P_{BB} denote the RF power consumption and baseband power consumption, respectively. The variables $\eta_{PA}, \sigma_{feeder}, \sigma_{DC}, \sigma_{MS}, \sigma_{cool}$ denote power amplifier efficiency, RF feeder losses, DC losses, MS losses and cooling losses, respectively.

4.3.2 Proposed H-C-RAN Energy Model

The above EARTH power consumption models cannot be directly used within the proposed C-RAN architecture for three reasons [77]. Firstly, the multiple BBUs reside in one cloud infrastructure where housing, power supply and cooling are shared, so the energy consumption of BBU per RRH should be reduced. Secondly, by virtualisation, the baseband computational resources can be dynamically allocated, and the BBU application can be run only when necessary. Finally, the energy consumption incurred by the fronthaul links between RRHs and the BS cloud are not at all considered by the EARTH model. The EARTH model cannot reflect the variations of computational resources, as a result, a new model under cloud-based cellular architectures is required. In order to properly evaluate the performance of the proposed framework, this thesis proposes a new energy usage. model where power

consumption of BBUs and the RRHs are calculated separately and modified to include cooling and fronthaul power consumption:

$$P_{hcran} = P_{radio} + P_{fronthaul} + P_{BScloud} \quad \textbf{(8)}$$

where P_{hcran} denotes total power consumption in the radio side (P_{radio}) plus fronthaul power consumption ($P_{fronthaul}$) plus power consumption in the BS cloud ($P_{BScloud}$).

4.3.2.1 Power Consumption at the Radio Side

The BBU is decoupled from the radio side and centralised in the BS cloud. The power consumption model in (7) is modified to come up with a generalised component based power consumption model of RRH_j within C-RAN, denoted as P_j, which is formulated as:

$$P_j = N_{TRX} \cdot \frac{\frac{\rho_j P_{max}}{\eta_{PA}} + P_{RF}}{(1-\sigma_{DC})(1-\sigma_{MS})} \quad (9)$$

It should be noted that feeder and cooling loses are negligible since the RRH is closer to the antennae and cooling for RRH is done by natural air. This model applies to both the MRRH and the PRRH. The total power consumption in the radio side of H-C-RAN then becomes:

$$P_{radio} = \sum_{j=1}^{N_R} P_j \quad (10)$$

The objective of the research in the radio side is to minimise P_{radio} in (10) which is formulated as:

$$\min \sum_{j=1}^{N_R} P_j \tag{11}$$

such that r_{min}^{k} in equation (2) is satisfied.

4.3.2.2 Power Consumption Model in the Fronthaul

The authors in [101] have shown that the relative effect of the fronthaul power consumption is non-negligible for scenarios with an increasing number of small (low power) BSs as in H-C-RAN. The fronthaul for C-RAN can either be fibre, microwave or copper connections or a mixture of these as described in Section 2.4.3. In this thesis, fibre fronthaul links are considered due to their high bandwidth and low latency characteristics. The RRHs are linked to the BS cloud using star topology due to its simplicity. The power contribution of the fibre fronthaul link is modelled as [102]:

$$P_{fronthaul} = \left\lceil \frac{1}{max_{dl}} \left(\sum_{a=1}^{m} N_a \right) \right\rceil P_s + \left(\sum_{a=1}^{m} N_a \right) P_{dl} \tag{12}$$

where, max_{dl} and m denote the maximum number of downlink interfaces available at one aggregation switch in the BS cloud and number of BS types used in the network respectively. The brackets $\lceil f(x) \rceil$ denotes the ceil function which takes as input a real number and gives as output the least integer that is greater than or equal to $f(x)$. The variables a and N_a denote the BS type and total number of RRHs regardless of their type, respectively. When $a = 1$, it means the RRH type is an MRRH and when $a = 2$, the RRH is a PRRH. The variables P_s and P_{dl} denote the power consumed by the aggregate switch and the power consumed by one downlink interface in the aggregation switch used to receive the downlink traffic, respectively.

4.3.2.3 Power Consumption at the Base Station Cloud

The total power consumption in the BS cloud, $P_{BScloud}$, consists of cooling power $P_{cooling}$, as well as the sum of power consumptions of all active GPPs:

$$P_{BScloud} = P_{cooling} + \sum_{i=1}^{N_G} P_{GPP_i} \tag{13}$$

where $P_{cooling}$ and P_{GPP_i} denote the power consumed by cooling and the power consumption of GPP_i, respectively. N_G denotes the total number of active GPPs to be minimised. The power consumption model of a standard GPP is as follows [82]:

$$P_{GPP_i} = P_0^{GPP_i} + \Delta_P^{GPP_i} . P_{max}^{GPP_i} . \rho_{GPP_i} \tag{14}$$

where $P_0^{GPP_i}$, $P_{max}^{GPP_i}$ and $\Delta_P^{GPP_i}$ denotes idle mode power consumption, i.e., 0% CPU utilisation, the maximum power consumption of the GPP at 100% CPU utilisation and the GPP power gradient, which is dependent on the type of GPP, respectively. The parameter ρ_{GPP_i} denotes the CPU utilisation of GPP_i. The assignment of vBBUs to GPPs is denoted by *VP* with association values $vp_{ij} \in [0,1]$ such that:

$$vp_{ij} = \begin{cases} 1 & \text{if } vBBU_j \text{ is assigned to } GPP_i \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

For a given assignment VP, the CPU utilisation of GPP_i can be calculated by:

$$\rho_{GPP_i} = \frac{\sum_{j=1}^{N_v} \rho_{vBBU_j} \bullet vp_{ij}}{C_{cap}^i} \tag{16}$$

where C_{cap}^{i} is the GPP maximum CPU capacity which is uniform for all GPPs. The RRH_j traffic is to be processed by a specific $vBBU_j$ as such it is necessary to calculate the amount of CPU for $vBBU_j$ which is described in Section 4.3.2.4 below.

4.3.2.4 Baseband Processing Power Model

The user traffic dynamics from RRH_j need to be converted to CPU processing resources i.e., the baseband workload denoted W_j in giga operations per seconds (GOPS). A model for converting user traffic dynamics from cell areas to baseband CPU processing power has been proposed in [103] which has been adopted in this thesis. This model states that the computing power in GOPS for a single user $k \in \boldsymbol{U}_j$ is calculated as:

$$W_k = \beta_{BB}(30Ant + 10Ant^2 + \frac{20 \bullet Mod \bullet D \bullet L}{6}) \frac{N_{j,k}}{50} \tag{17}$$

$$\boldsymbol{W}_j = \rho_{vBBU_j} = \sum_{k \in U_k} \boldsymbol{W}_k \tag{18}$$

where β_{BB} is the percentage of the baseband functions moved to the BS cloud, Ant is the number of antennas used per user, Mod is the modulation bits, D is the coding rate used, L is the number of MIMO layers used, $N_{j,k}$ is the number of currently used PRBs by user k. W_j denotes the total baseband workload in terms of CPU for RRH_j.

The key objective is to find an assignment VP that minimises total GPP power consumption:

$$\min_{VP_{ij}} \sum_{i=1}^{N_G} P_{GPP_i} \tag{19}$$

Subject to:

$$\sum_{i=1}^{N_G} vp_{ij} = 1 \tag{20}$$

$$\sum_{j=1}^{N_v} vp_{ij} * \rho_{GPP_i} \leq C_{cap}^{i} \tag{21}$$

Constraint (20) means one vBBU can only be assigned and should be assigned to one GPP. Constraints (21) requires that the total CPU resources of vBBUs assigned to a GPP cannot exceed the GPP maximum capacity C_{cap}^{i}. It should be noted that other constraints like memory, disk and bandwidth resources of vBBUs of a GPP should be less than the maximum capacity. In this thesis only CPU resource is considered as a vBBU placement constraint since the power consumption of a GPP scale linearly with the CPU.

4.4 Proposed RRH Sleeping Algorithm

In this section, the details of the proposed BS switching off algorithm is presented, which aims to minimise the power consumption at the radio side (P_{radio}) shown in equation (11). Only the PRRH are to be switched off whereas all the MRRH remain switched on to maintain coverage. The PRRH switch off algorithm is centralised and located at the BS cloud in the GCC which can guarantee a considerable energy saving performance. Figure 4-3 shows the proposed PRRH switching off framework block diagram consisting of various modules as described below.

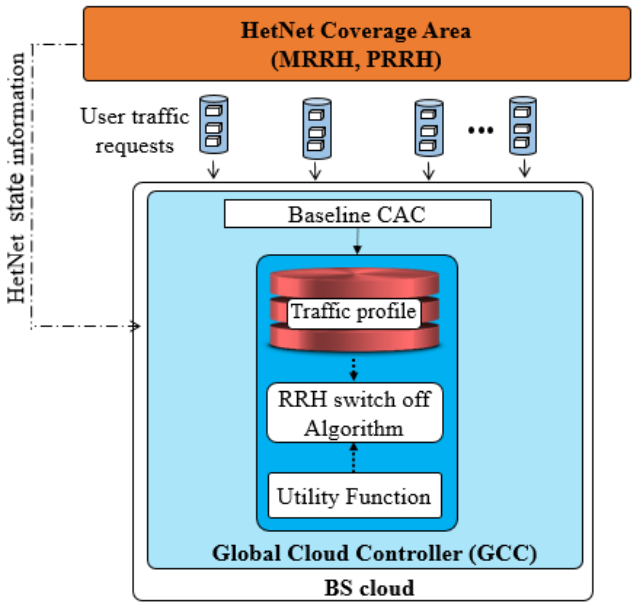

Figure 4-3. PRRH switching framework.

- ***HetNet Coverage Area:*** The RRHs cover cells where users are located. The user requests from the coverage area are sent to the GCC in the BS cloud where they are processed. The requests pass through a baseline CAC in the GCC. The user traffic requests states some QoS related information like the number of PRBs requested, the CQI which states the modulation and coding rate in the downlink and so on.
- ***Global Cloud Controller (GCC):*** The GCC is a centralised controller that run the PRRH switch off algorithm and is the main module located in the BS cloud. The GCC receives user traffic requests and HetNet state information (RRH traffic load, interference status). Various submodules of the GCC are described below:
    - ***Call Admission Control (CAC):*** The baseline CAC simply drops connection requests in the BS cloud when a certain threshold of the number of users is met.

- ***Traffic Profile***: The traffic profile is stored in a database in the GCC. This is the traffic profile introduced in Section 2.3.3 in Figure 2-6.
- ***Utility Function:*** This is a function introduced in Section 4.4.1 which is used for testing if the PRRH can be switched off based on the PRRH's traffic load, current datarate, interference level and power consumption.
- ***RRH Switch off Algorithm:*** The PRRH switch off algorithm will be explained in Section 4.4.2 and is also located in the GCC. After collecting the HetNet state information and calculating the utility functions of each PRRH, the PRRH switch off algorithm is then invoked.

Since the traffic load fluctuates over time, the number of active PRRHs should track this fluctuation to make a trade-off between satisfying the user service requirement and saving power. As soon as the users joins the network and are associated with their respective RRHs, the utility of each PRRH can then be calculated as described in Section 4.2.3. The utility function is used as a performance metric to determine if a certain PRRH should be switched off or kept on. After calculating the utility function, the proposed PRRH switch OFF algorithm is implemented. The proposed PRRH switch off algorithm runs periodically. The traffic period $T_{total} = 24$ hours is divided into Y equal time intervals of T as shown in Figure 4-4. At each time spot $t(i), i = 1, \ldots, Y$, the GCC carries out the proposed PRRH switching OFF algorithm to determine all PRRHs working modes based on the current network information. First the

utility function used for PRRH switching off will be presented followed by the PRRH switch off algorithm.

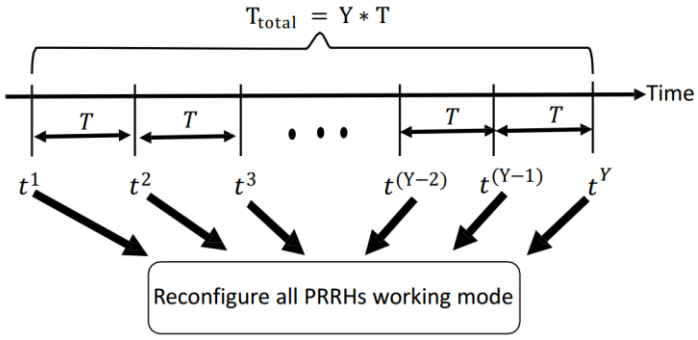

Figure 4-4. PRRH switch off over time.

4.4.1 Utility Function Calculation

The utility function is only defined for PRRHs as they are the ones to be tested for switching off. The traffic load of PRRHs has more significant fluctuations in space and time due to a number of factors such as user mobility and behaviour, as well as the fact that each PRRH supports fewer simultaneous UEs. Therefore, when designing the rule of determining which PRRH should first be tested to switch off, more factors should be included in the rule, instead of just the PRRH traffic load as widely used in literature. After having UEs in the HetNets associated with their respective RRHs and performing resource allocation, a utility function for each PRRH can then be computed. The proposed utility function depends on the total datarate of served UEs, the PRRH traffic load, as well as the power consumption of the PRRH and the received interference signal strength from the nearby cells. The utility function in [76] is adopted and modified to include energy consumption ratio of the PRRH ($\frac{P_x}{P_{max}}$), as follows:

$$U_x = \frac{\varepsilon \frac{R_x}{R_{max}} + \beta \frac{\rho_x}{\rho_{max}}}{\frac{I_x}{I_{max}} * \frac{P_x}{P_{max}}} \quad \text{where,} \tag{22}$$

$$\rho_x = \frac{\text{PRBs used by PRRH UEs}}{\text{Total PRBs available to the PRRH}} \tag{23}$$

where U_x denotes the utility function of $PRRH_x$ and P_x is the power consumption of $PRRH_x$. The variables ε and β are the weighting coefficients of the considered factors and satisfy $\varepsilon + \beta$ = 1. The variables R_x, ρ_x, P_x, denote the total datarate of all UEs serviced in $PRRH_x$, the traffic load on $PRRH_x$ and power consumption of $PRRH_x$. The term I_x is the received interference signal strength from the nearby RRHs. Specifically, I_x is once calculated over all PRBs assuming that all PRBs of nearby RRHs are used to account for the worst interference case and then stored for the following utility calculation. The four terms with subscript max are the maximum values of the corresponding terms respectively and are used for normalisation. The above mentioned parameters needed to calculate the utility can be obtained from the network information collected by the GCC in the BS cloud.

In (22), the utility value is monotonically increasing with the two terms on the numerator, when the sleeping algorithm tries to switch off PRRHs, the load of the cell has to be considered (i.e., PRB occupation ratio and the number of served UEs) and the throughput of the cell. The aim is to switch off those PRRHs with relatively lower traffic load, and thus trade-off between energy saving and network capacity. Besides, the interference term on the denominator is used to give more priority to those PRRHs which receive

relatively strong interference signal from nearby MRRHs and other PRRHs to be switched off. Also the power consumption term in the denominator gives priority to those PRRHs with high power consumption to be switched off.

4.4.2 PRRH Switch OFF Algorithm

The PRRH switch OFF algorithm is shown in Algorithm 1. The algorithm runs in the GCC at the BS cloud and takes as input the HetNet state information (RRHs traffic load and RRHs interference status). The output is the number of PRRHs that are switched off. In line 1, a set of already switched off PRRHs are stored in the set R_{off} whereas a set of active PRRHs are denoted as R_{on}. In line 2, the total number of switched off PRRHs is initialised to zero as there are no PRRHs switched off in the beginning. In line 3, the utility function U_x of each PRRH is calculated based on the state information collected in the GCC. In line 4, the PRRHs from R_{on} are sorted in ascending order by their utility function values into set R_{sort} such that the first $PRRH_1$ in R_{sort} is of the worst performance whereas $PRRH_{N_2}$ in R_{sort} is of the best performance. Then in line 5, the iterative value of PRRHs denoted x is initialised to 1 to begin with the first PRRH in R_{sort}. In line 6, the users of $PRRH_x$ where $x = 1$ are transferred into nearby PRRHs or MRRHs. Then if all the users of $PRRH_x$ (line 7) are transferred, update R_{off} (line 8) and R_{on} (line 9).Then switch off $PRRH_x$ and update the system UE association information (line 10). If the users in $PRRH_x$ cannot be transferred to other RRHs, keep $PRRH_x$ switched on (line 10) and try the next PRRH which is $PRRH_{x+1}$ by incrementing the value of x (line13). Then repeat from line 4 to line 14 until all the PRRH have been checked (until$x = N_2$) for switching off.

Algorithm 1 PRRH switch off algorithm

Input: HetNet sate information, user association.

Output: Number of switched off PRRHs.

1: Initialise a set of switched off PRRHs as R_{off} and a set of switched on PRRHs as R_{on}.

2: Initialise the number of switch OFF PRRHs as $N_{off} = 0$.

3: Calculate utility values for each PRRH based on HetNet state information collected by the GCC.

4: Sort PRRHs by their increasing utility values into set $R_{sort} = [PRRH_x : x = 1, \dots, N_2]$ where $PRRH_1$ and $PRRH_{N_2}$ are the worst and the best performing PRRHs, respectively.

5: Set the PRRH incremental value $x = 1$.

6: Transfer UEs of $PRRH_x$ with the least utility to the nearby RRHs.

7: If (all UEs of $PRRH_x$ are transferred) Then

8: $R_{off} = R_{off} + \{PRRH_x\}$

9: $R_{on} = R_{on} - \{PRRH_x\}$

10: Switch off $PRRH_x$ and update user association information.

11: Else

12: Keep $PRRH_x$ switch on.

13: Increment x and go to step 6.

14: End If

15: Repeat from step 4 to line 14 until all the PRRH have been checked (until $x = N_2$).

This subsection has described in detail how power consumption in the radio side (P_{radio}) can be minimised by switching off underutilised PRRHs in the H-C-RAN framework. The following subsection describes how power consumption can be minimised in the BS cloud using cloud computing baseband workload consolidation.

4.5 Baseband Processing Workload Consolidation Framework

This section presents the solution for the optimisation problem in (19) by minimising energy consumption at the BS cloud using cloud computing based

workload consolidation mechanism. Workload consolidation is a cloud computing mechanism for processing workload into fewer number of computing servers or GPPs to save energy by switching off underutilised servers. Workload in the context of C-RAN means baseband CPU processing power and is measured in GOPS. Workload consolidation is achieved through optimised virtual machine placement which is the process of mapping VMs or vBBUs to GPPs by ensuring to improve power efficiency and resource utilisation. The vBBUs must be distributed in an efficient way such that no system or a request starves for the response from BS cloud. If vBBUs are allowed to migrate from one GPP to another, then it shows that it is a dynamic placement. If there are no migrations between systems, then it is said to be a static placement of vBBUs. In our case, dynamic placement of the vBBUs is addressed. Figure 4-5 shows the difference between a normal data centre and server consolidation data centre. Efficient management of host machine increases efficiency of the data centre and improves service response time.

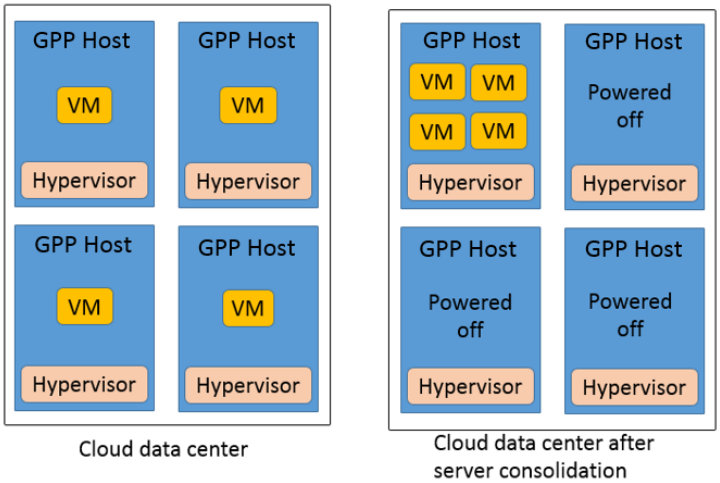

Figure 4-5. Cloud server consolidation.

As shown in Figure 4-5, if the workload of the host system is balanced then the system may respond quickly to their users and improve the overall system

efficiency. The main aim is to pack RRH user traffic load into fewer number of GPPs by minimising the value of N_G at a reasonable time frame which maintains QoS. Finding an optimal solution for vBBU placement optimisation problems in (19) can be an incredibly difficult task. The problem is formulated as a bin-packing problem, which has been proved to be NP-hard. The following section explains the heuristic bin packing algorithms like next-fit, first-fit and first-fit decreasing that have been proposed to minimise the number of GPPs in the BS cloud to save energy. Moreover, two heuristic algorithms based on simulated annealing and genetic algorithm are also proposed for minimising the number of GPPs in the BS cloud.

The H-C-RAN workload consolidation model is shown in Figure 4-6. The model is an improved version of the proposed RRH sleeping framework proposed previously in Section 4.4. The model consists of various functional modules as described below.

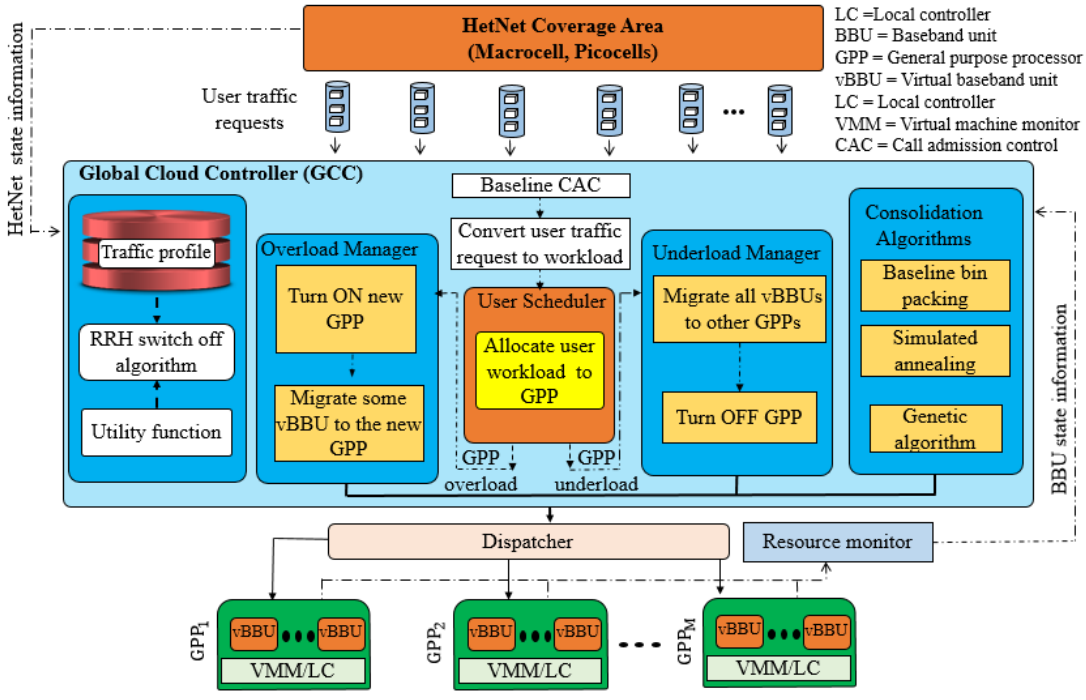

Figure 4-6. H-C-RAN Workload consolidation model.

Some of these functional modules have been described in Section 4.4. The new entities added to support workload consolidation are as follows:

- ***GCC:*** To support the new functionality required, the GCC is modified and consists of various modules as described below.
  - *User traffic request to CPU workload converter:* The user request is converted to CPU workload W_k as calculated using (17).
  - *User scheduler:* Users are scheduled to GPPs based on CPU resources. The user workload is allocated to GPPs such that overloading and underloading are avoided as described below.
  - *Overload manager:* Overloading is when GPP utilisation is above maximum threshold. Overloading of GPPs normally happens when traffic load increases requiring more GPPs to be turned on. During GPP overloading, some vBBUs are migrated to other GPPs.
  - *Underload manager:* Underloading occurs when CPU utilisation is below the minimum threshold. If traffic load in the coverage area is low, the processing workload will also be reduced. In such conditions, the underutilised GPPs are then turned off to save energy by issuing a shutdown GPP command.
  - *Consolidation algorithms module*: This is where the conventional bin packing algorithms, the GA and the SA simulation algorithms reside. The network operator chooses the algorithm to use and the chosen algorithm is invoked every T time for a period of 24 hours.
- ***Workload dispatcher/aggregate switch*:** The dispatcher receives requests from the GCC on where to route the data from RRH to GPP. User

data is then directed from RRH to a specific GPP without passing through GCC.

- ***Virtual machine monitor (VMM)/Local controller (LC) and Monitor*:** The VMM/LC is located in each GPP and it creates and runs vBBUs in the GPP. The VMM, also called the hypervisor, regularly collects CPU utilisation status from the GPPs and forwards to the monitor module which then sends status feedback to the GCC to check for overloaded or underloaded GPPs.

4.5.1 vBBU Live Migration among GPPs

The vBBU live migration (VLM) is the lossless movement of a vBBU state from a source GPP to a target GPP which will then take over all the services provided by the source GPP as shown in Figure 4-7. However, it is important that this transfer occurs in a manner that balances the requirements of minimising migration time which is the duration between when migration is initiated and when the original vBBU may be finally discarded. A vBBU is really made up of two basic components: The vBBU's storage called virtual hard disk (VHD) and the vBBU's configuration or state. Often, a VHD is located on a shared storage area network (SAN), and the vBBU configuration or state is what's running in a host GPP's processor and memory. With VLM, the vBBU's state and configuration is copied from one physical host GPP to another, but the VHD is not moved. VLM is invoked when a GPP is overloaded or when a GPP is underloaded. VLM is also necessary when the source host GPP need to be taken down for maintenance, upgrade and repair. The proposed VLM procedure falls into three successive phases: migration preparation, migration execution and migration completion as shown in Figure 4-8.

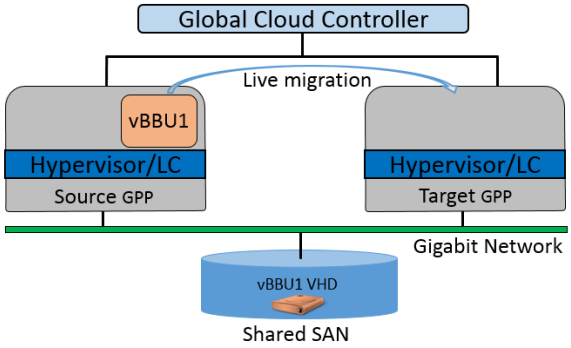

Figure 4-7. An illustration of H-C-RAN vBBU live migration.

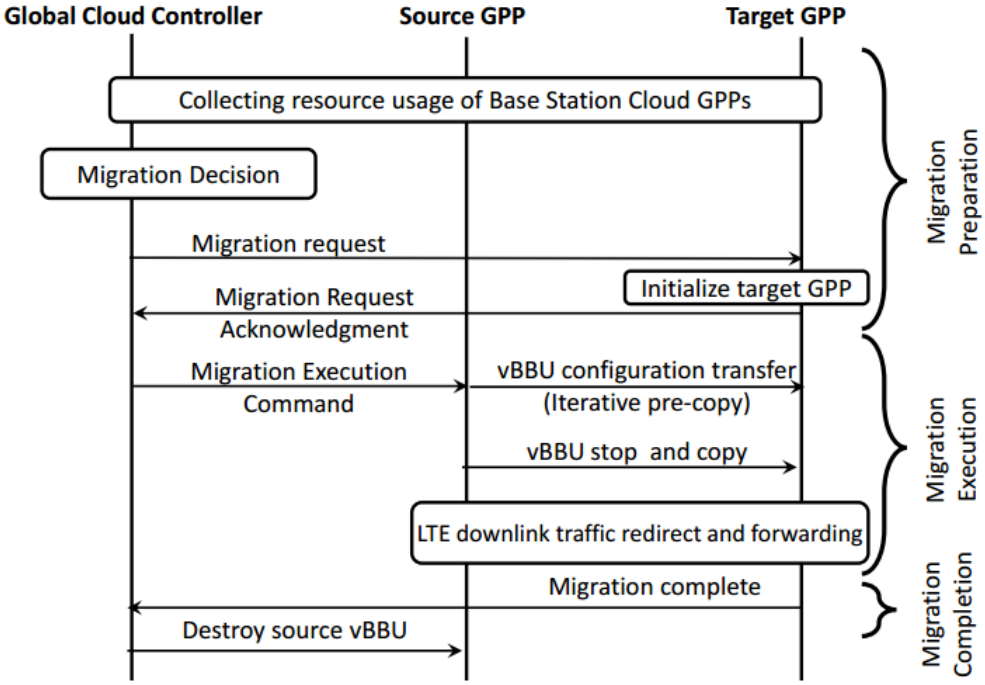

Figure 4-8. H-C-RAN vBBU live migration flow chart.

1. ***Migration preparation:*** After a migration decision has been made by the GCC on which vBBU to migrate, a migration request is sent to a source GPP and a target GPP. It is initially confirmed that the necessary processing resources are available on the target GPP and reserve a VM container of that vBBU size. After the target GPP has been initialised with reserved resources, an acknowledgement request is sent to the GCC.

2. ***Migration execution***: During execution phase, a migration execution command is sent to the source GPP to begin the migration. The state or configuration of the vBBU is copied to the target GPP. This is done using the memory iterative pre-copy technique where first the whole memory pages of the vBBU are copied to the target GPP including modified/dirty pages while the vBBU is still actively running on the source GPP. During stop and copy procedure, the vBBU will be stopped on the source GPP and will be resumed on the target host. An address resolution protocol (ARP) is then issued to redirect and forward the LTE downlink traffic to target GPP. The new vBBU then restores the context information of the user and the vBBU operate as normal.
3. ***Migration completion***: The target GPP then issues a migration complete command to the GCC which then issues a command for destroying the VM of the vBBU in the source GPP.

4.5.2 Baseline vBBU Placement Schemes

This section will describe the baseline vBBU placement algorithms that are derived from conventional bin packing schemes used for minimising the number of GPPs in the BS cloud. The classical bin packing problem consists of packing a series of items with sizes in the interval (0, C_{cap}] into a minimum number of bins with capacity, C_{cap}. The vBBU placement problem can be modelled as a bin packing problem where the aim is to pack user traffic items from cell areas into a set of GPPs such that the number of servers used are minimised and hence the power consumption reduction. Four conventional bin packing schemes, namely NF, FF and FFD and FBP have been adopted to define the baseline vBBU placement algorithms. These algorithms will help

evaluate the baseline performance of standard vBBU placement algorithms against which comparisons can be made when the proposed more sophisticated heuristic based methods are used.

4.5.2.1 Next-Fit (NF) Algorithm

The NF algorithm works as follows: Initially all bins are empty and it starts with bin $b = 1$ and item $e = 1$. If bin b has residual capacity for item e, assign item e to bin b and consider item $e + 1$. Otherwise consider bin $b + 1$ and item e. Repeat until all items are assigned. The NF algorithm never considers bins again that have been left behind. Thus, the wasted capacity therein leaves room for improvement. The NF algorithm is shown in Figure 4-9 where the bin size is 1 and the item queue start from the far right from 0.5, 0.8, 0.9, 0.2, 0.4 and 0.1. The items are placed on the GPPs with the first GPP being GPP_1.

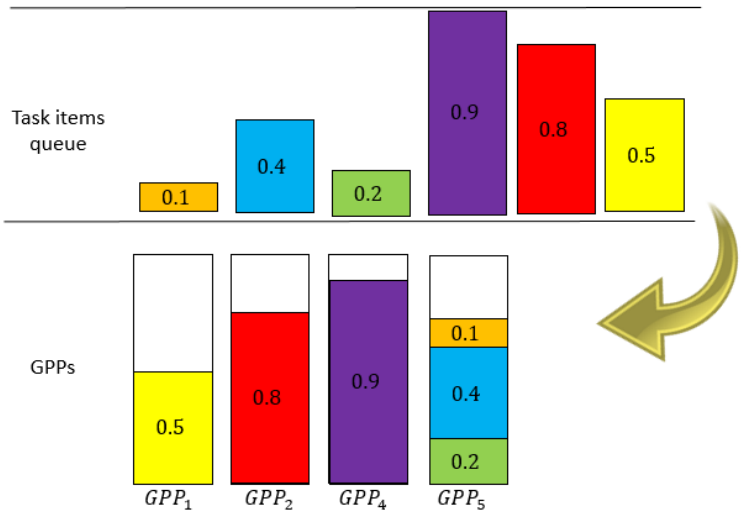

Figure 4-9. Next fit algorithm.

4.5.2.2 First-Fit (FF) Algorithm

In the FF algorithm, initially all bins are empty and it start with the current number of bins $N_b = 0$ and item $e = 1$. Consider all bins $\boldsymbol{B} = [b: b = 1, \ldots, N_b]$ and place item e in the first bin that has sufficient residual capacity.

If there is no such bin, increment b and repeat until all items are assigned as shown in Figure 4-10 where the bin size is 1 and the item queue start from the far right.

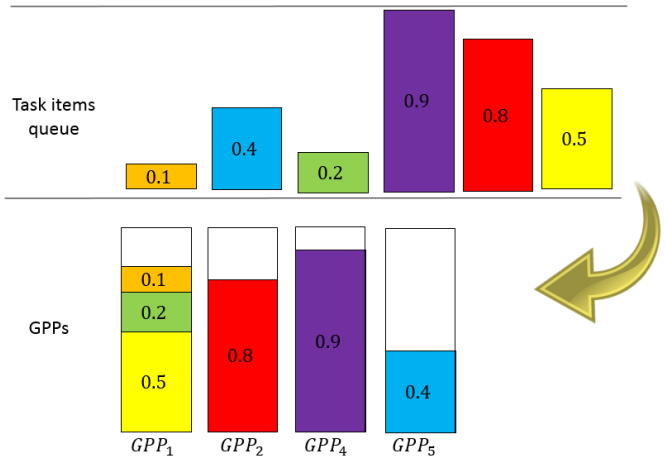

Figure 4-10. First fit algorithm.

4.5.2.3 First-Fit Decreasing (FFD) Algorithm

The FFD algorithm is shown in Figure 4-11 where the bin size is 1 and the item queue start from the far right.

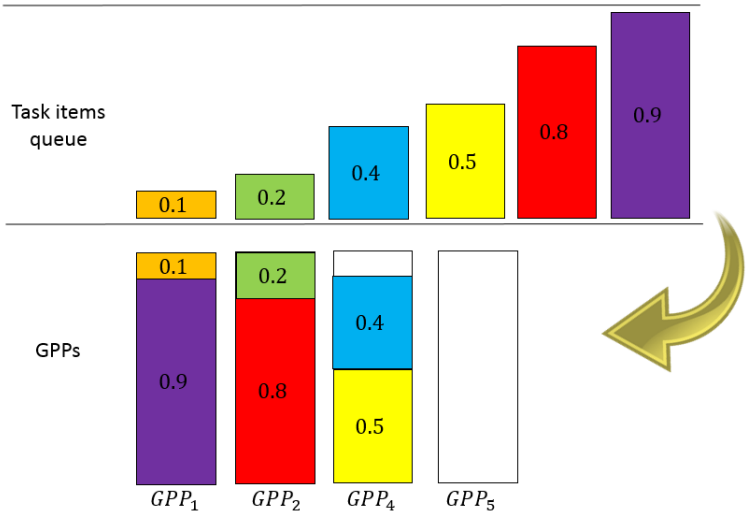

Figure 4-11. First fit decreasing algorithm.

In the FFD algorithm, initially all bins are empty and all items are arranged in increasing order. Initially start with bin $b = 1$ with item $e = 1$. If bin b has

residual capacity for item e, assign item e to bin b, and consider item $e + 1$. Otherwise consider bin $b + 1$ and item e. Repeat until all items are assigned.

4.5.3 Proposed Heuristic vBBU Placement Algorithms

In this section, two approximation heuristic algorithms called simulated annealing and genetic algorithm based vBBU placement algorithms are proposed to solve the vBBU bin packing problem and dynamically minimise the number of GPPs.

4.5.3.1 Simulated Annealing based vBBU Placement Algorithm

Simulated Annealing (SA) first proposed by Kirkpatrick *et al*. in 1983 is a general purpose combinatorial optimisation and probabilistic technique for approximating the global optimum of a given function in a large search space [104]. SA algorithm was originally inspired from the process of annealing in metal work which involves heating and cooling the metal to alter its physical properties due to the changes in its internal structure. As the metal cools its new structure becomes fixed, consequently causing the metal to retain its newly obtained properties. In SA, a temperature variable is kept to simulate this heating process. The temperature is initially set to high and then allowed to slowly cool as the algorithm runs. While this temperature variable is high the algorithm will be allowed, with more frequency, to accept solutions that are worse than the current solution. This gives the algorithm the ability to jump out of any local optimums it finds itself in early on in execution. As the temperature is reduced so is the chance of accepting worse solutions, therefore allowing the algorithm to gradually focus in on an area of the search space in which a close to optimum solution can be found. This gradual 'cooling' process is what makes the simulated annealing algorithm remarkably effective at finding a

close to optimum solution when dealing with large problems which contain numerous local optimums. SA is chosen due to its excellent nature at avoiding being stuck at local optimums.

There are basically four components in the SA algorithm: (i) problem configuration; (ii) initial and new configuration; (iii) decision of acceptance; and (iv) temperature scheduling.

1. ***Problem configuration:*** The assignment of vBBUs $\boldsymbol{V}$ to GPPs $\boldsymbol{M}$ is the configuration. Making use of constraint that one vBBU can only be assigned to one GPP, and to reduce the number of variables and the searching space, an integer array is used to represent the assignment:

$$V_{conf} = \left[vBBU_1, vBBU_2, \dots, vBBU_j, \dots, vBBU_{Nv}\right]$$

   The index of the array is the vBBU number, and the value indexed by the vBBU number in the array is the GPP number to which the vBBU is assigned. For example, if there are 5 vBBUs indexed from 1 to 5 and 3 GPPs numbered from 1 to 3. $V_{conf} = [\,1, 2, 1, 3, 2]$ means that $vBBU_1$, $vBBU_3$ are assigned to GPP_1; $vBBU_2$ and $vBBU_5$ are assigned to GPP_2; $vBBU_4$ is assigned to GPP_3.

2. ***Initial configuration and new configuration generation:*** The algorithm initially allocates vBBUs to GPPs randomly and this allocation serves as the starting point of the annealing process and will be referred to as the initial state of the system. At every evolution step, the configuration needs to be changed into a new configuration state, so the configurations change slowly moving toward better configurations. To obtain a new state from the current state, the system is perturbed using one of the two methods below:

- relocation of a single randomly selected vBBU from the GPP to which it is currently allocated to a randomly selected GPP,
- randomly selecting two vBBUs currently allocated to two different GPPs and exchanging their positions.

3. ***Decision of acceptance:*** While generating new configurations provides the candidate new state, the acceptance criteria decide which configuration becomes the new state. First a check is performed to determine if the new solution's energy consumption $E_{new} = \sum_{i=1}^{N_G} P_{GPP_i}(VP')$ is lower than the current solution $E_{curr} = \sum_{i=1}^{N_G} P_{GPP_i}(VP)$ where VP' is the vBBU placement of the new configuration. If the energy difference $\delta E = E_{new} - E_{curr} \leq 0$, the new state is accepted without any qualification. If $\delta E \geq 0$, then the new state is accepted with a probability, $P_r(\delta E)$ calculated as [105]:

$$P_r(\delta E) = e^{-\delta E/Temp} \tag{24}$$

where $Temp$ is the scheduling temperature which is described in the following section. In the proposed algorithm, if a new configuration is feasible and has lower energy consumption than the previous state, it will be accepted as the new state. If a solution is found to be infeasible, then the probability of acceptance is used to either reject or accept the new solution.

4. ***Temperature scheduling:*** The algorithm starts initially with $Temp$ set to a high value, and then it is decreased at each step following some annealing schedule, which may be specified by the user, but must end with $Temp = 0$ towards the end of the allotted time budget. In this thesis, a temperature decrement function of the form $F(Temp) = \xi * Temp$ will be used where

ξ lies in the interval [0.9, 1] as in [105]. At high temperatures, $e^{-\frac{\delta E}{Temp}} \cong 1$, state changes involving relocation of items with large sizes are likely to be accepted. At low temperatures, $e^{-\frac{\delta E}{Temp}} \cong 0$, generated configurations are likely to be accepted only when they have a smaller energy.

The proposed algorithm for SA is shown in Algorithm 2. The algorithm takes as inputs the GPP CPU capacities and the CPU requirements for vBBUs and outputs the assignment of vBBUs to GPPs (VP) together with the minimised value of N_G.

Algorithm 2 Simulated Annealing Algorithm

Input:	CPU load of GPPs, ρ_{GPP_i} and vBBU CPU requirements, ρ_{vBBU_j}.
Output:	vBBU to GPP map, VP' and minimum number of GPPs, N_G.
1:	$Temp = T_{max}$
2:	Generate initial random placement VP.
3:	While $Temp > 0$ do
4:	Calculate the energy of VP as $E_{curr} = \sum_{i=1}^{N_G} P_{GPP_i}(VP)$
5:	For $i = 1\ to\ T_{max} * N_v$ do
6:	Generate new placement VP' using perturbation.
7:	Calculate the energy of VP' as $E_{new} = \sum_{i=1}^{N_G} P_{GPP_i}(VP')$
8:	Calculate the energy difference, $\delta E = E_{new} - E_{curr}$
9:	If $\delta E \leq 0$ then
10:	Set placement VP' as new state
11:	Else
12:	Generate a random number $\eta\ \epsilon\ \{0,1\}$
13:	If $\eta < e^{-\delta E/Temp}$ then
14:	Set placement VP' as new state
15:	End If
16:	End If
17:	End For
18	Temp = $\varepsilon * Temp$ where $\varepsilon\ \epsilon\ [0.9,1]$
19:	End while

First, the temperature is initialised to a high value T_{max} (line 1). Then an initial placement VP is generated randomly (line 2). As the temperature value

remains greater than zero (line 3), the energy consumed for the current VP configuration is calculated as $E_{curr} = \sum_{i=1}^{N_G} P_{GPP_i}(VP)$ (line 4). Then a loop is performed for $T_{max} * N_G$ times (line 5), and during this times a new VP' configurations is generated (line 6). The energy of the new generated configuration VP' is then calculated as $E_{new} = \sum_{i=1}^{N_G} P_{GPP_i}(VP')$ (line 7). Then the temperature difference is calculated (line 8) to check whether to accept or reject the configuration VP' (line 9 to line 15) and try the next configuration if VP' consumes more energy. Decrement the value of temperature with the temperature decrement function $F(Temp) = \xi * Temp$.

4.5.3.2 Genetic Algorithm based vBBU Placement Algorithm

A Genetic Algorithm (GA) based vBBU placement algorithm is also proposed that aims to minimise the energy consumption and number of active GPPs within the BS cloud. GA is a method for solving both constrained and unconstrained optimisation problems based on a natural selection process that mimics biological evolution and it is great for finding solutions to complex search problems. In GA, a population of candidate solutions to an optimisation problem is evolved toward better solutions. Each candidate solution has a set of properties (its chromosomes or genotype) which can be mutated and altered. GA is often used in fields such as engineering to create incredibly high quality products thanks to their ability to search through a huge combination of parameters to find the best match. In the following section, the GA steps will be presented.

4.5.3.2.1 *Encoding*

In the proposed GA based vBBU placement scheme, the representation of chromosomes and genes are designed with bins/GPPs in mind. A gene represents a single bin/GPP which is a group of vBBUs as shown in Figure 4-12. A chromosome consists of N_v number of genes. The chromosome is represented by a group of genes which is a collection of bins/GPPs as shown in Figure 4-13.

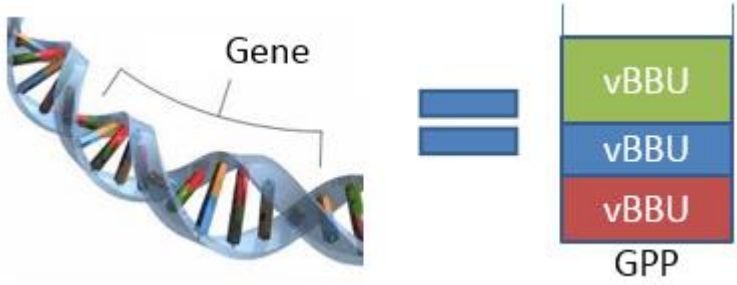

Figure 4-12. Gene representation.

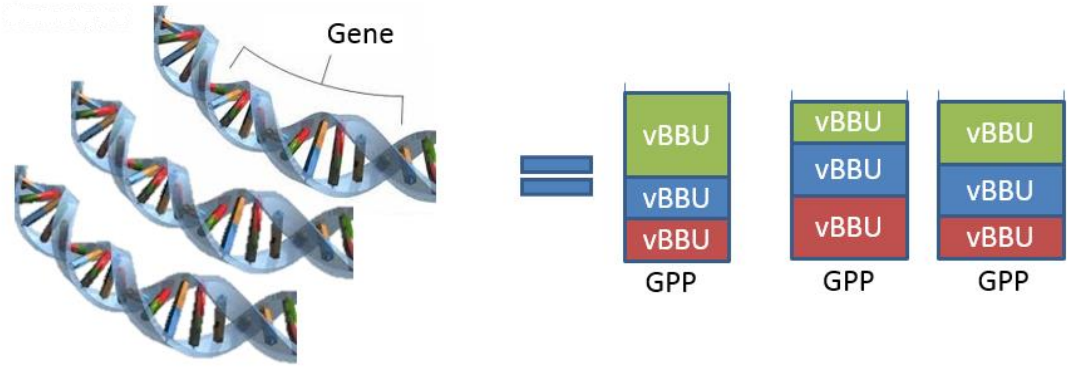

Figure 4-13. Chromosome representation.

4.5.3.2.2 *Initialisation:*

The population size depends on the nature of the problem, but typically contains several hundreds or thousands of possible solutions. Often, the initial population is generated randomly, allowing the entire range of possible

solutions. This population is usually randomly generated and can be of any desired size, from only a few individuals to thousands.

4.5.3.2.3 *Fitness Evaluation*

Each member of the population is then evaluated and 'fitness' for that individual is calculated. The fitness value is calculated by how well it fits with the desired requirements. The most obvious fitness function for this problem is the number of items used by the solution, but it does not create smooth search space that genetic algorithm can explore. To make the search space smooth, a function that takes the fill of bins in the solution into account is used and it looks like this [106]:

$$F = \frac{\sum_{i=1}^{N_G} \left(\frac{f_i}{C_{cap}^i}\right)^q}{N_G} \tag{25}$$

where F is the fitness of the solution, N_G is the number of bins, f_i is the fill of the i^{th} bin which is the summation of the vBBU CPU capacities, C_{cap}^i is the capacity of the bin and q is a constant greater than one. The q constant controls whether the more filled bins are preferred. Larger values should be used in case more filled bins are preferred. The GA should minimise the fitness function where the vBBUs are packed in as small number of bins as possible.

4.5.3.2.4 *Selection*

There is need to be constantly improving the overall fitness of the population. During each successive generation, a proportion of the existing population is selected to breed a new generation. Individual solutions are selected through a fitness-based process, where fitter solutions (as measured by a fitness function) are typically more likely to be selected. Selection helps us to do this

by discarding the bad designs and only keeping the best individuals in the population. During selection, two parent's chromosomes (genotypes) are randomly selected.

4.5.3.2.5 *Genetic Operators*

The next step is to generate a second generation population of solutions from those selected through a combination of genetic operators: crossover and mutation.

Crossover: In this method, to create a child chromosome, two parents are involved. Combining the vector entries of two different elements to form a child. Once when a new generation is formed, it replaces the existing generation. The hope is that by combining certain traits from two or more individuals it will create an even 'fitter' offspring which will inherit the best traits from each of it's parents. The crossover operation is described in Algorithm 3.

Algorithm 3 Cross Over Operation

Input:	Two parent chromosomes, $C^s = c_1^s c_2^s, \dots, c_z^s$ and $C^q = c_1^q c_2^q, \dots, c_z^q$
Output:	Single child chromosome, $C^v = c_1^v c_2^v, \dots, c_z^v$
1:	$F^s = fitness(C^s)$
2:	$F^q = fitness(C^q)$.
3:	Select a random gene c_{rand}^s in C^s
4:	Select a random vBBU $rand_{vBBU}^s$ in c_{rand}^s
5:	Select a random gene c_{rand}^q in C^q
6:	Select a random vBBU $rand_{vBBU}^q$ in c_{rand}^q
7:	Exchange $rand_{vBBU}^s$ with $rand_{vBBU}^q$ in $c_{rand}^s \in C^s$ to from C^v
8:	Output C^v

Mutation: There is a need to add a little bit of randomness into the population's genetics otherwise every combination of solutions created would be in the

initial population. Mutation is a way of giving birth to a child where only a single parent is involved. Mutation typically works by making very small changes at random to an individual's genome. The mutation operator involves randomly picking up a gene in the chromosome and then reversing its value. Algorithm 4 shows how the mutation operation works.

Algorithm 4 Mutation Operation Algorithm

Input:	Single chromosomes, $C = c_1 c_2, \dots, c_z$
Output:	Single mutated child chromosome, $C' = c_1' c_2', \dots, c_z'$
1:	Select a random gene c_{rand} in C
2:	Select a random vBBU $rand_{vBBU}$ in c_{rand}
3:	Select a random gene c'_{rand} in C'
4:	Select a random vBBU $rand_{vBBU}$, in c'_{rand}
5:	Exchange $c'_{rand} = c_{rand}$ in C'
6:	Output C'

4.5.3.2.6 *Algorithm structure*

The GA algorithm is shown in Algorithm 5 outlining the flow of the GA algorithm that has already been explained.

Algorithm 5 Genetic Algorithm

Input:	CPU load of GPPs, ρ_{GPP_i} and vBBU CPU requirements, ρ_{vBBU_j}.
Output:	vBBU to GPP map, VP' and minimum number of GPPs, N_G.
1:	$Generation = 0$.
2:	Generate a population of individuals Pop.
3:	Select the best individual I_b in Pop.
4:	While $Generation \leq maximum_{generation}$ do

5: For $each\ individual\ I_b\ \epsilon\ Pop$ do
6: Calculate its fitness function.
7: Use the roulette selection to select another individual based on CPU utilisation to pair.
8: End for
9: For $each\ pair\ \epsilon\ parents$ do
10: Call crossover operation.
11: End for
12: For $each\ individual\ I_b\ \epsilon\ Pop$ do
13: Call mutation operation.
14: End for
15: Find the best candidate $P_{best}\ \in Pop$.
16: If $P_{best}\ is\ better$ Then
17: $I_b = P_{best}$
18: End If
19: Generation = Generation +1
20: End while
21: Output $P_{best} = vBBU\ allocation\ map\ VP$
22: Output N_G

4.6 Advanced Call Admission Control using Fuzzy Logic

The previous two sections, Section 4.4 and Section 4.5, have described ways of minimising energy consumption in the radio side and in the BS cloud respectively using a baseline CAC mechanism which simply drops the requests when there are not enough resources in the BS cloud. Such CAC scheme is not suitable for next generation wireless networks like 5G where a plethora of services exist and chances of traffic congestion are very high. The 5G cellular networks will need to cope with the explosive increase of traffic requests from these devices to avoid network overload and traffic congestion.

In this thesis, a novel Fuzzy-logic based CAC scheme with pre-emption specifically designed for 5G C-RAN architecture is proposed. In this scheme,

cloud bursting technique is used during congestions, where some delay tolerant low-priority Non Real Time (NRT) connections are pre-empted and outsourced to a public cloud with a penalty charge to allow the acceptance of high priority Real Time (RT) connections. CAC is an RRM scheme that offers an effective way of avoiding network congestion and plays a key role in the provision of guaranteed QoS in the RAN. The basic function of a CAC algorithm is to accurately decide whether a connection can be accepted into a resource-constrained network without violating the service commitments made to the already admitted connections. A good CAC scheme aims to optimise call blocking probability, resource utilisation and energy efficiency.

The proposed Fuzzy based CAC scheme replaces the baseline CAC schemes used in Section 4.4 and Section 4.5. Fuzzy-logic scheme is used for performing CAC in 5G C-RAN because of its simplicity and robustness [6]. Fuzzy-logic techniques resembles the human decision making with an ability to generate precise solutions from certain or approximate information. Fuzzy-logic avoids uncertainties and computational complexities brought by many CAC schemes and does not require precise inputs, and can process any number of inputs. Fuzzy-logic incorporates a simple, rule based approach based on natural language to solve problem rather than attempting to model a system mathematically. Figure 4-14 shows the proposed fuzzy CAC system model diagram for 5G C-RAN which is located in the BS cloud inside the GCC.

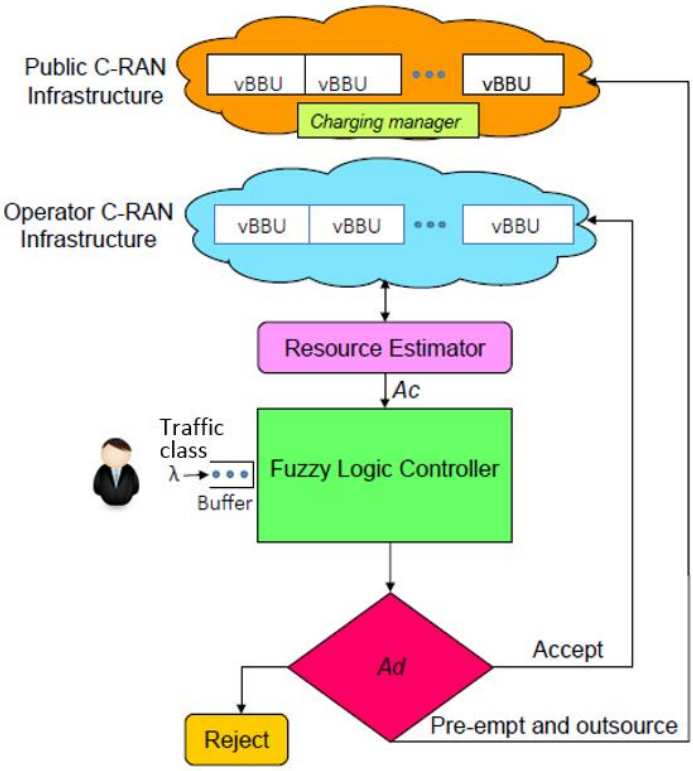

Figure 4-14. The model of the fuzzy logic based CAC with pre-emption for C-RAN 5G.

The model consists of various modules consisting of the operator's C-RAN infrastructure for normal processing of requests when the congestion is low and a third party public C-RAN infrastructure for handling requests for the operator's C-RAN during congestion. Connection requests that are processed in the public infrastructure are charged a certain price by the charging manager depending on the type of service and the size of the connection request.

The resource estimator estimates the available capacity in the operator's C-RAN infrastructure and indicates whether the cloud is congested or not. The model also consists of the fuzzy controller which performs the CAC decisions for incoming requests from users. The fuzzy controller takes as inputs two variables which are the service type St and the normalised available capacity, Ac. The output is the admittance decision, Ad, which is either, accept a

request, reject a request or pre-empt some low priority requests and outsource them to a public cloud. The traffic requests are divided into two groups as shown below;

- RT class - These are called guaranteed bit rate (GBR) and include VoIP, live streaming, video call and real time gaming. This type of services are delay sensitive.
- NRT class - These are called variable bit rate (VBR) and include buffered streaming, and transmission control protocol (TCP) based services like web browsing, email, file transfer protocol (ftp), and point to point (p2p). This type of services are delay tolerant to some extent.

4.6.1 Cloud Bursting Technique for Pre-empted Connections

The cloud bursting technique allows the operators to dynamically extend their infrastructure by renting third-party resources [16]. During congestion of the operator's C-RAN infrastructure, when a high priority RT connection arrives as illustrated in Figure 4-15, and the cloud is congested, two things happen, either the low priority NRT connections are pre-empted from the operator's C-RAN and then moved into the public C-RAN infrastructure to accommodate the high priority RT connections or the RT connection is dropped if there are no NRT connections to pre-empt in the operator's C-RAN.

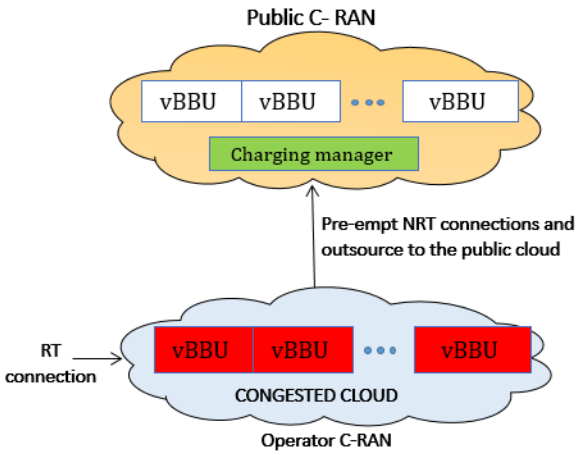

Figure 4-15. Cloud bursting model for pre-empted connections during arrival of RT connections.

RT connections are never outsourced to the public cloud because they are delay sensitive. Only NRT connections are outsourced to the public cloud. An agreement is made between the operator and the public cloud operator and a certain price is charged for outsourcing some NRT connections. When a NRT connection arrives and the operator's cloud is congested, the NRT connection is forwarded to the public cloud as shown in Figure 4-16 with a certain price penalty where the request will be charged by the charging manager.

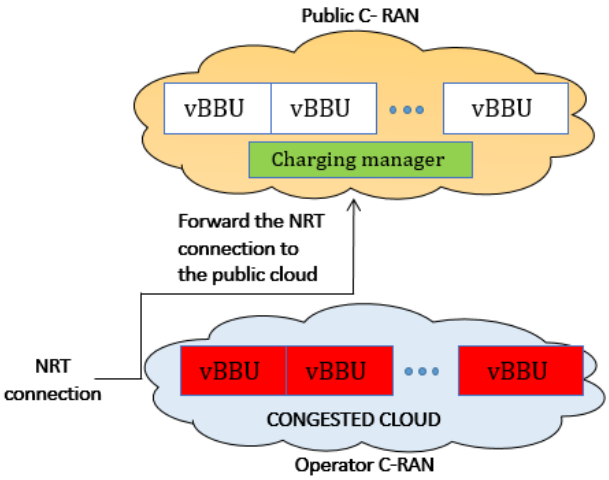

Figure 4-16. Cloud bursting model for new arrival of NRT connections.

4.6.2 Structure of Fuzzy Logic Controller

The fuzzy controller of the proposed scheme takes two inputs: (i) available capacity, Ac; and (ii) service type, St and outputs the admittance decision, Ad. Below is the description of the structure of the proposed fuzzy-logic controller.

4.6.2.1 Membership Functions

The trapezoidal and triangular membership functions are chosen for the proposed algorithm. The membership functions for input and output linguistic parameters are shown in Figure 4-17. The values of the membership functions have been chosen based on commonly used values of membership functions in various literature. For the fuzzy controller, the term sets for Ec, St, Ac and Ad are defined as follows:

i) T(St) = {NRT, RT}

ii) T(Ac) = {NotEnough, Enough}

iii) T(Ad) = {Accept, Reject, Preempt}

4.6.2.2 Fuzzy Rule Base

The fuzzy rule base consists of a series of fuzzy rules, shown in Table 4. These control rules are of the following form: IF ’condition’, THEN ’action’. Example: if St is ’RT’ and ’Ac’ is ’Not Enough’ then Preempt.

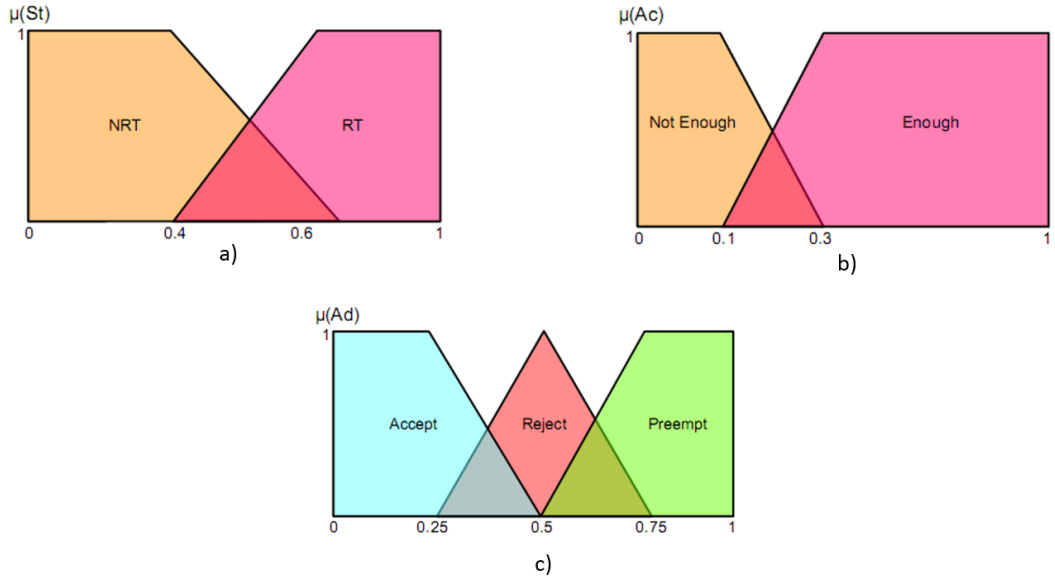

Figure 4-17. Membership functions for (a) Service type, St (b) Available capacity, Ac and (c) Admittance, Ad.

Table 4. Fuzzy Rule Base for Fuzzy Controller.

Rule	$\boldsymbol{St}$	$\boldsymbol{Ac}$	$\boldsymbol{Ad}$
1	RT	Not Enough	Preempt
2	RT	Not Enough	Preempt
3	RT	Not Enough	Preempt
4	RT	Enough	Accept
5	RT	Enough	Accept
6	RT	Enough	Accept
7	NRT	Not Enough	Preempt
8	NRT	Not Enough	Preempt
9	NRT	Not Enough	Preempt
10	NRT	Enough	Accept
11	NRT	Enough	Accept
12	NRT	Enough	Accept

4.6.3 Defuzzification Method

The Centre of Gravity (COG) [1] method is used for defuzzification to convert the degrees of membership of output linguistic variables into crisp/numerical values. The COG method is adopted since the membership functions used are simple triangular and trapezoidal shapes with low computational complexity and can be expressed as [1]:

$$D_{COG} = \frac{\int_{z} \mu(z) z dz}{\int_{z} \mu(z)\, dz} \tag{26}$$

4.7 Saving Energy in the Mobile Device

So far, energy has been saved in the radio side and BS cloud in H-C-RAN and advanced CAC schemes have been implemented in the BS cloud to improve resource utilisation and allocation to facilitate efficient energy usage by the network.

This section presents the framework for saving energy in the mobile devices within the proposed 5G C-RAN architecture. In this thesis, an energy efficient MEC scheme for 5G H-C-RAN (EMCC) is proposed. In EMCC, an application from the mobile device is partitioned into tasks where some tasks are executed on the local device whereas others are offloaded and executed in parallel in VMs in the BBU pool while taking into consideration the transmission cost and delay.

4.7.1 Communication Model

Since mobile device will connect to the MRRH or the PRRH, the system bandwidth B is divided into N_{ch} channels or PRBs as stated in Section 4.2.

Denote a set of channels in the system as $\boldsymbol{C} = [c: c = 1,2, \dots, N_{ch}]$. It is considered that each mobile device user runs an application, which can be split into several tasks. Each task T_k of user device k can be executed either locally on the mobile device or remotely on the MEC server by computation offloading. Consider that the user device k can offload T_k either via the MRRH or via the PRRH. Denote $a_{k,m,c} = [0,1]$ as the offloading decision profile of user k where $m = [1,2,3]$ is the user device chosen modes which are computing locally, transmitting via the MRRH and transmitting via the PRRH, respectively. $a_{k,m,c} = 1$ means device k uses mode m to offload task T_k through channel c and if $a_{k,m,c} = 0$, otherwise. The item c is meaningless when $m = 1$ as there are no channels in local computing, thus $a_{k,1,1} = 1$ is taken as the indicator that device k select local computation. In case the mobile device k offloads the task T_k via the MRRH on channel c, the uplink data rate of mobile user can be computed as

$$r_{k,c}^{M} = w \log_2 (1 + \frac{P_k^{\,M} H_k^{\,M}}{\sum_{j=1, j \neq k}^{N_{UE}} a_{k,2,c} P_j^{\,S} H_j^{\,M} + \sigma^2}) \tag{27}$$

where w is the channel bandwidth, P_k^M is the transmission power from user k to the MRRH, H_k^M is the channel gain between user k and the MRRH. The denominator in (27) is the interference caused by other mobile users using the same channel for transmission. The variable σ^2 denotes the background noise power. The total uplink data rate of mobile user k to the MRRH is calculated as:

$$r_k^M = \sum_{c=1}^{N_{ch}} a_{k,2,c} r_{k,c}^M \tag{28}$$

Similarly, if the mobile user offloads a task via the PRRH through channel c, the uplink data rate is given as

$$r_{k,c}^S = w \log(1 + \frac{P_k{}^S H_k{}^S}{\sum_{j=1, j\neq k}^{N_{UE}} a_{k,3,c} P_j{}^M H_j{}^S + \sigma^2}) \tag{29}$$

$$\text{and, } r_k^S = \sum_{c=1}^{N_{ch}} a_{k,3,c} r_{k,c}^S \tag{30}$$

4.7.2 Mobile Application Computation Model

Each task of mobile user k is denoted as $T_k = (B_k, D_k, t_k^{\max})$. Here B_k denotes the size of computation input data in bytes (e.g., the program codes and input parameters) involved in the computation task T_k and D_k denotes the processing requirement in million instructions per second (MIPS) required to accomplish the computation task T_k. The variable t_k^{max} denotes the maximum latency required by the computation task T_k or the execution deadline in milliseconds (ms).

4.7.2.1 Local Computation

Local computation is when the mobile device k executes its computation task T_k locally on the mobile device. Denote F_k^l as the computation capability of the mobile device user k in MIPS. It is assumed that mobile devices can have various computation capabilities. The execution time for executing task T_k for user k can be expressed as:

$$t_k^l = \frac{D_k}{F_k^l} \tag{31}$$

The energy consumed by the mobile device user k for local computation can be expressed as:

$$e_k^l = t_k^l P_a \tag{32}$$

where P_a is the power consumed by the device when active. It is assumed that only the mobile application in consideration is running in the mobile device. Even when more than one applications are running in the mobile device, the proposed scheme will still offload the mobile application in consideration to the MEC server and reduce energy consumption.

4.7.2.2 Remote Computation

When a mobile device chooses computing its task in the BS cloud, the input data can be transmitted to the VM in the BS cloud through the MRRH or the PRRH. The mobile device user would incur the extra overhead in terms of time and energy for transmitting the computation input data to the BS cloud via wireless access. In case device k offloads T_k via MRRH, the total time duration (t_k^M) can be calculated as transmission time plus time during remote execution of task T_k:

$$t_k^M = \frac{B_k}{r_k^M} + \frac{D_k}{F_k^{GPP}} \tag{33}$$

where F_k^{GPP} is the computation ability of the allocated VM in the GPP server. The total energy consumed by the mobile device via offloading through the MRRH can then be calculated as:

$$e_k^M = \frac{B_k}{r_k^M} P_k^M + \frac{D_k}{F_k^{GPP}} P_{idle} \tag{34}$$

where P_{idle} is power consumption of the mobile device when inactive. Similarly for offloading via the PRRH,

$$t_k^S = \frac{B_k}{r_k^S} + \frac{D_k}{F_k^{GPP}} \tag{35}$$

$$and, e_k^S = \frac{B_k}{r_k^S} P_k^S + \frac{D_k}{F_k^{GPP}} P_{idle} \tag{36}$$

4.7.3 Problem Formulation

The aim is to minimise energy consumption in the mobile device by offloading some application tasks to the BS cloud. The optimisation problem can be formulated as:

$$\min_{\{a_{k,m,c}\}} \sum_{k=1}^{N_{UE}} (a_{k,1,1} e_k^l + \alpha_{k,2} (P_k^M \frac{B_k}{r_k^M} \sum_{c=1}^{N_{ch}} a_{k,2,c} + P_{idle} \frac{D_k}{F_k^{GPP}}) \tag{37}$$

$$+\alpha_3 (P_k^S \frac{B_k}{r_k^S} \sum_{c=1}^{N_{ch}} a_{k,3,c} + P_{idle} \frac{D_k}{F_k^{GPP}}))$$

$$\text{such that, } C1: a_{k,1,1} t_k^l \le t_k^{\max}, k \in \boldsymbol{U} \tag{38}$$

$$C2: \sum_{k=1}^{N_{Nu}} \sum_{c=1}^{N_{ch}} a_{k,m,c} \le N_{ch}, \quad m = \{2,3\} \tag{39}$$

$$C3: \sum_{c=1}^{N_{ch}} a_{k,1,c} . \sum_{c=1}^{N_{ch}} a_{k,2,c} = 0, \quad k \in \boldsymbol{U}, c \in \boldsymbol{C} \tag{40}$$

$$C4: \sum_{c=1}^{N_{ch}} a_{k,2,c} . \sum_{c=1}^{N_{ch}} a_{k,3,c} = 0, \quad k \in \boldsymbol{U}, \mathrm{c} \in \boldsymbol{C} \tag{41}$$

$$C5: \sum_{c=1}^{N_{ch}} a_{k,1,c} . \sum_{c=1}^{N_{ch}} a_{k,3,c} = 0, \quad k \in \boldsymbol{U}, \mathrm{c} \in \boldsymbol{C} \tag{42}$$

$$C6: \sum_{k=1}^{N_{UE}} a_{k,m,c} = 1, \quad k \in \boldsymbol{U}, \mathrm{m} \in \boldsymbol{M}, \mathrm{c} \in \boldsymbol{C} \tag{43}$$

where $\alpha_{k,m} = 1(\sum_{k=1}^{N_{UE}} a_{k,m,c} > 0,\ j = [2,3])$.The function $1(h($ is an indicator function which is equal to 1 when h is true and zero otherwise. The first constraint C1 ensures that the delay constraints are met. The second constraint ensures that the total number of channels allocated to mobile devices does not exceed the total number of channels in the system. Constraint (C3 to C5) state that for each device k, only one offloading model can be implemented. Constraint C6 states that only one channel can be allocated to only one mobile device.

4.7.4 Proposed EMCC Model

The EMCC architecture is shown in Figure 4-18. Parallel execution of tasks in EMCC is done by deploying the tasks of an application on various VMs in the BBU pool. In the BS cloud, GPPs host VMs which host $cloneApps$ responsible for executing the mobile device tasks. Several $cloneApps$ can exist to execute the mobile application tasks in parallel and results sent back to the mobile device. The proposed framework model of EMCC is shown in Figure 4-19.

The system modules are described as follows:

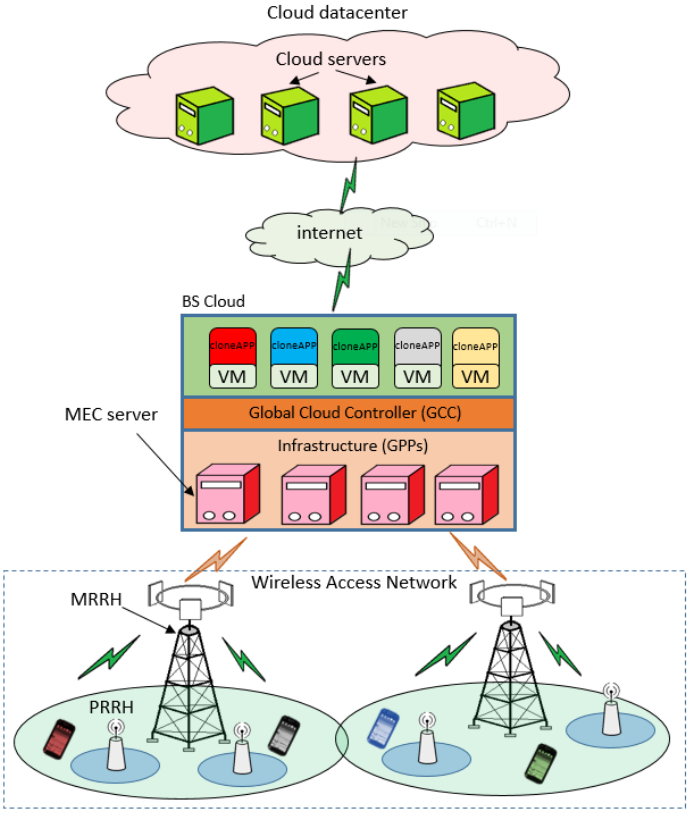

Figure 4-18. Proposed EMCC architecture.

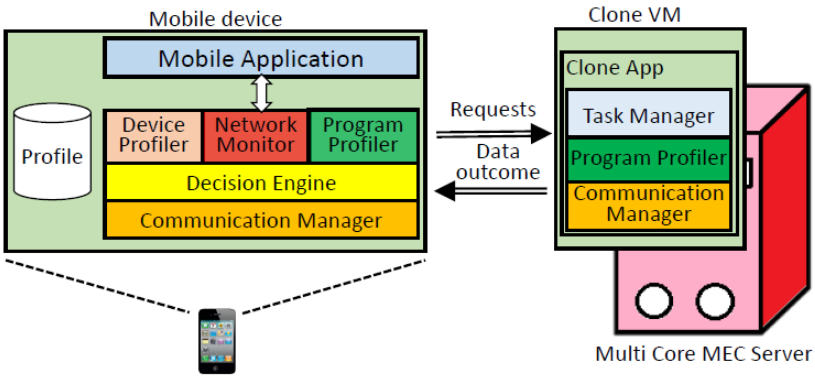

Figure 4-19. Proposed EMCC framework model.

- **Device profiler:** Collects mobile device hardware context at runtime and passes the information to the decision engine. The profile context includes the battery SoC, average CPU utilisation and memory usage.

- **Network monitor**: Collects network related context at runtime and passes the information to the decision engine. The context includes cell connection state, bandwidth and cell signal strength.
- **Decision engine**: Consists of a set of cost estimation models like the delay or execution time and the energy models. Based on the received context above, the decision engine decides on when, where and how to offload the task. The flow chart in Figure 4-20 shows operation of the decision engine.

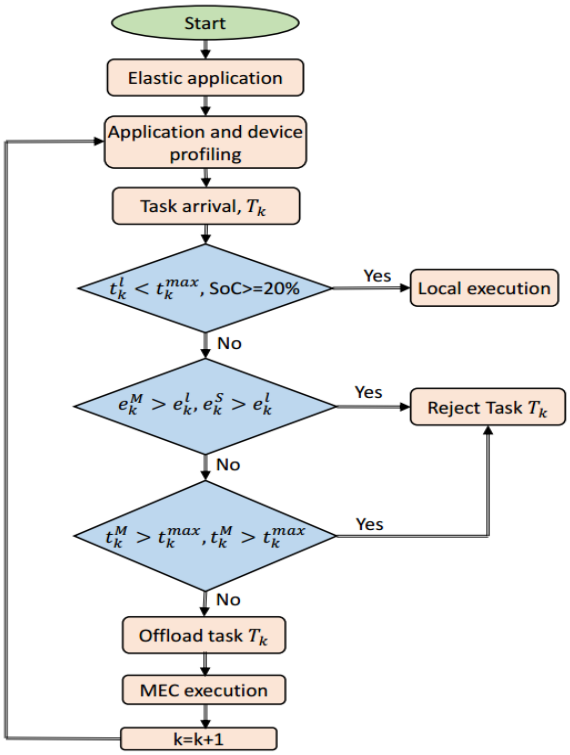

Figure 4-20. EMCC decision engine flow chart

The decision engine starts with the arrival of a task, T_k, if local execution time is less than the maximum delay tolerable and the battery SoC ≥ 20%, the task is executed locally. Else if the offloading delay deadlines are met and energy is saved using offloading, the task is offloaded to the VM in the MEC server. The next task $T_k + 1$ then follows the same order in the flow chart.

- **Program profiler:** Runs on both the mobile device and the clone VM. The program profiler tracks the execution of the program and collects program context information such as total instructions executed, execution time, memory allocated. The profile is updated at every invocation and it is stored in the mobile device database.
- **Communication manager**: This runs in both the mobile device and in the clone VM of the application. It creates and maintains connection between the client and the server side. It serialises the code on the client side and deserialises the request from the client at the cloud side. It also keep the client and the server VM in sync. The communication manager checks if the required files and programs exist in the server side, else it contacts the client device to fetch the files and related libraries for remote execution.
- **Task manager**: On the cloud side, the task manager handles the requests and executes the offloaded code on the VM after state synchronisation between the mobile device and the clone VM.

4.8 Concluding Remarks

This chapter has presented the proposed energy efficient framework for saving energy in 5G HetNet C-RAN while maintaining the required QoS. The proposed framework has been divided into four parts which are;

i) Saving energy at the radio side of C-RAN using BS sleeping mechanism using baseline CAC where the PRRHs are switched off based on a utility function that depends on the traffic load, the network interference and BS power consumption. Only the PRRHs are switched off whereas the MRRHs are kept active to maintain coverage and to avoid signal outage,

ii) Saving energy at the BS cloud using baseband workload consolidation where the baseband processing workload from the radio side is packed into fewer number of GPP servers according to the traffic load such that underutilised servers are switched off to save energy while maintaining QoS. The underutilised GPPs to be shut off have their vBBUs migrated to other GPPs.

iii) Using advance CAC mechanism in the BS cloud to further improve performance. A novel Fuzzy logic based CAC scheme with pre-emption in 5G C-RAN is proposed. In this scheme, cloud bursting technique is used during congestions where some delay tolerant low-priority NRT connections are pre-empted and outsourced to a public cloud to allow the acceptance of high priority RT connections,

iv) Saving energy in the mobile device by proposing a MEC paradigm called EMCC where an application from the mobile device is partitioned into tasks where some tasks are executed on the local device whereas others are offloaded and executed in parallel in high processing MEC servers in the BS cloud while taking into consideration the transmission cost and delay.

5 Simulation Framework and Results

5.1 Introduction

Cellular communication systems are very complex in nature due to the architectures and the environments in which they are deployed. For example, the architecture of a wireless cellular system needs to be designed in such a way as to provide high data rates while satisfying constraints such as transmit power, bandwidth and minimum QoS requirements. The system often involves a large number of random events over time and space such as the location and the number of simultaneous mobile users, propagation conditions, interference and power level settings, and user traffic demand. This combination of system complexity and challenging variable environments leads to design and analysis of problems that are not analytically tractable by applying traditional, non-simulation based techniques [107]. Prototypes are often not available and testing different candidate features in the field would often prove to be too expensive and time consuming. Moreover, simulations of wireless systems are essential in the research and standardisation phases whenever a new technology is under development. Consequently, the most convenient way to evaluate the performance of the system is by computer simulations, which have become a widely adopted methodology [108]. This simulation-based approach not only reduces the high cost of actually implementing a real system, but also saves a significant development time.

Over the last two decades, the development of powerful and inexpensive computers together with robust software packages has grown rapidly. As a

consequence, computer aided design and analysis techniques are readily available, and are usually referred to as simulations. An important motivation for adopting a simulation approach in this thesis is that it provides a valuable insight into system behaviour and performance before considering expensive real implementation. For this reason, the aim of this chapter is to present the evaluation guidelines specifying both the simulation methodology and tools used in assessing the overall performance of the proposed resource management 5G C-RAN framework and also to present and discuss simulation results.

5.2 Performance Metrics

The following performance metrics will be used for evaluating the performance of the proposed H-C-RAN framework.

1. Power Consumption: This is the total power consumed in the entire network measured in watts that considers the power consumption both from the BS cloud and radio side. Power consumption metric is important in showing the network that will have less OPEX and also environmentally friendly with minimal CO_2 emissions. Equation (8) is used to compute the total power consumption of C-RAN.
2. Throughput: The data rate or throughput of a mobile device is defined as the number of information bits per second successfully delivered or received and is an important performance metric in terms of QoS. According to the classical Shannon formula, the attainable user throughput (bits/second) between RRH_j and UE k are given by r_{UE}^{k}. In the simulation, the value of r_{UE}^{k} was calculated by using LTE CQI (Channel Quality

Indicator) where the user reports the CQI index stating the modulation and the coding scheme of the user, then each CQI index correspond to a transport block size (TBS) index determined how many bits can be transmitted per TTI (1 ms). Then the value of the obtained bits is multiplied by 1000 to get bps. Consequently, the overall system throughput of RRH_j is given by:

$$R_j = \sum_{k \in \mathrm{U}} r_{UE}^{k} \tag{44}$$

3. Energy Efficiency: Energy efficiency is a key performance indicator in this thesis as it reflects the network energy (or equivalent power) consumed relative to the system capacity. Energy efficiency (η_{EE}) in bits per joule is the ratio of the achievable throughput (bits/second) to the total network power consumption, which can be derived as [12]:

$$\eta_{EE} = \frac{\sum_{j \in \mathrm{R}}^{N_R} R_j}{\text{Network Power Consumption}} \tag{45}$$

4. Area Power Consumption (APC): This is defined as the ratio of the average total consumed power to the corresponding network area measured in watts per km^2:

$$P_A = \frac{P_{network}}{A} \ \ \text{watts/km}^2 \tag{46}$$

where A is the total coverage area in the radio side and $P_{network}$ is the power consumption within the network.

5. Call Blocking Probability: The call blocking probability in the simulation within an RRH was calculated as follows[109]:

$$P_{block} = \frac{\text{Number of blocked calls}}{\text{Total number of calls}} \quad (47)$$

When a cellular system is designed, the blocking probability should be less than a particular predetermined value, P_{th} = 5% [12].

6. Resource Utilisation: It is the ratio of the processing workload from cells to the maximum capacity of servers utilised in the network. It shows how efficiently resources are being utilised.
7. Number of active BBUs: This is the total number of BBU servers used in the entire network according to traffic load. The C-RAN is expected to use less BBUs due to the workload consolidation mechanism. The number of servers in C-RAN is the output of running the workload consolidation algorithm.
8. Statistical Multiplexing Gain (θ): The ratio of infrastructure (GPP servers) used in traditional LTE system to the infrastructure used in C-RAN. The higher the value the better the gain. Thus,

$$\theta = \frac{\text{\#active BBU servers in traditional LTE-A system}}{\text{\#active BBU servers in C-RAN}} \quad (48)$$

5.3 Simulation Platform

Over the past years, a variety of software packages have been developed, which have been widely used to simulate communication systems. Simulation models can be built using a general-purpose programming language such as C/C++, Java or FORTRAN and writing the appropriate code, or by using a

graphical model builder such as SIMULINK or OPNET. The graphical model builders are relatively simple to use by clicking and dropping functional blocks on the computer screen and linking them together to create a simulation model in a hierarchical block diagram form. Models can have a number of representations ranging from floating-point sub-routines to bit-level implementations of sub-routine models.

As an alternative to using a graphical block diagram editor for model building, one could use an intermediate (pseudo) language such as the MATLAB command language, which is one of the popular numerical computing environments and programming languages. MATLAB was chosen as the main system modelling tool for this thesis to demonstrate original concepts, for problem solving, and for rigorous comparison with the baseline model. There are a number of persuasive reasons for adopting MATLAB. Firstly, it combines excellent computational capabilities with easy-to-use graphical capabilities. It also contains a rich library of pre-programmed functions (m-files) for generating, analysing, processing and displaying signals together with specially developed add-on libraries (toolboxes) for communications and signal processing. It is also easy for MATLAB users to generate new m-files for user-dependent applications. Additionally, MATLAB code is very concise, making it possible to express complex digital signal-processing (simulation) algorithms using relative few lines of code and although MATLAB is relatively slow compared to the basic C/C++ or Java, running the MATLAB codes in powerful computers can overcome these slow execution speeds. For this

thesis, the computer specifications and MATLAB version used are summarised below.

Table 5. Simulation platform specifications

MATLAB version	PC specification	
MATLAB R2014a	OS	Windows 8 (64-bit)
	Processor	Intel Core i5-3470 CPU @3.20GHz
	RAM	8GB
	Hard Disk	500GB

5.4 Simulation Settings

To analyse the performance of the proposed H-C-RAN framework, a 2-tier simulation HetNet layout of a 19 MRRH overlaid by up to 12 PRRH per MRRH is considered. The bandwidth of 20MHz is considered and there are up to 3600 users in total with up to 8 users randomly generated within PRRH coverage (the PRRH has small coverage) and up to 100 users (because there a maximum of 100 PRBs allocated) for the MRRH. The number of users within the MRRH and PRRH follows the traffic profile in Figure 2-6. Each user is allocated one PRB per transmission time interval (TTI). The GPPs in the BS cloud consists of industry standard (ISS) blade servers called Intel Xeon processor E5540 [110] [111]. The E5540 is quad core (4 CPUs) with 45 GOPS per CPU hence a capacity of 180 GOPS. The servers are always operating at maximum efficiency (CPU always at maximum). Various parameter settings are shown in the tables below in Table 6, Table 7 and Table 8.

Table 6. Fronthaul and aggregate switch/ dispatcher settings [102]

Parameters	**Settings**
Maximum aggregate switch interfaces, max_{dl}	247

Number of MRRHs, N_1	19
Number of PRRHs, N_2	228
Number of BS types, m	2
Aggregate switch power consumption, P_s	300Watts
Aggregate switch interface power consumption, P_{dl}	1Watt

Table 7. Radio side parameter settings

Parameters	MRRH settings	PRRH settings
Cellular layout	19 cells	12 cells
System bandwidth, B	20MHz	20MHz
Number of PRBs	100	100
Shadowing standard deviation	8dB	10dB
Antenna pattern	3D	Omni
Antenna gain, $G_{k,n}^{j}$	14dBi	5dBi
Thermal noise density, σ^2	-174dBm/Hz	-174dBm/Hz
Transmission power ,P_{max}	46dBm	30dBm
UE distribution radius	500m	40m
Number of users	Up to 100	Up to 8
BS idle power [36], P_0	324 Watts	9 Watts
BS gradient slope [36], Δ_p	2.8	2.8
No. of user antennas, Ant	1,2,4,8	1,2,4,8
Modulation scheme, Mod	4QAM, 16QAM, 64QAM	4QAM, 16QAM, 64QAM
Coding rate, D	1/3 to 1	1/3 to 1
MIMO Layers, L	2	2
PA efficiency, η_{PA}	31.1%	6.7%
$\delta_{DC}, \delta_{MS}, \delta_{cool}$ [36]	6%, 7%, 9%	6%, 7%, 0%
BS antennas, N_{TRX}	1	1
$R_{max}, \rho_{max}, I_{max}$	100Mbps,100%,18dB	

Table 8. BS cloud parameter settings

Parameters	Settings
BS cloud cooling power [112], $P_{cooling}$	500W
GPP model [111]	Xeon Processor E5540
GPP capacity, C_{cap}^{i}	180 GOPS
GPP CPU lower threshold	30% of 180 GOPS = 54 GOPS
GPP CPU upper threshold	90% of 180 GOPS = 162 GOPS
GPP idle power [77], $P_0^{GPP_i}$	120Watts
GPP maximum power [77], $P_{max}^{GPP_i}$	215Watts
Server power gradient [77], $\Delta_P^{GPP_i}$	0.44
Total number of vBBUs, N_V	247

For the proposed GA based vBBU placement algorithm, the following parameters were used. In each generation, roulette selection is used to select two parents that will produce 50 child chromosomes. For each parent two rounds of roulette wheel selection is performed and the parent with better fitness is selected. Crossover probability is 100%, so each offspring is produced by crossover operation and with none being cloned. The probability of mutation used in this implementation is 50% because mutation is performed on 25 of 50 individuals. The value of q in the fitness function is set to 10 to give preference to more filled bins. The values of the items and bin capacity are normalised such that the bin capacity is one in relation to the 180GOPS bin capacity size and item sizes are greater than zero and less than one.

Table 9. GA simulation parameters.

Parameters	Settings
Population size	100 chromosomes
Mutation size	2 genes
Mutation probability	50%
Crossover probability	100%
Number of offspring's to produce	50
Selection type	Roulette wheel

Selection rounds	2
Selection size	2
Stop criterion type	Fitness change
Number of generations	100 generations

The table below shows simulation parameters for the MEC simulation.

Table 10. Parameters for the MEC simulation.

Task type	**Value**
MEC server processing, $\boldsymbol{F_k^{GPP}}$	44800MIPS
Mobile device processing, $\boldsymbol{F_k^l}$	1000-1200MIPS
Mobile device active power, $\boldsymbol{P_a}$	0.9 Watts
Mobile device idle power, $\boldsymbol{P_{idle}}$	1.3 Watts
Mobile transmission power, $\boldsymbol{P_k^s and\ P_k^M}$.	0.3 Watts

Table 11 shows the traffic classes for 4G from [113] that were used to generate user traffic in the coverage area within users.

Table 11. Traffic classes

***QCI**	**Service**	**Type**	**Delay**	***MBR**
1	VoIP	GBR	100ms	12Kbps
3	Conversational video	GBR	150ms	240Kbps
8	ftp	Non-GBR	300ms	512Kbps
9	www	Non-GBR	300ms	512Kbps

*QCI = Quality Channel Indicator, *MBR= Maximum Bit Rate.

Other relevant parameters are shown in Table 12.

Table 12. Other relevant settings

Parameters	**Settings**
Total number of users, N_{UE}	3600
Number of RRHs, N_R	247
Blocking probability threshold, P_{th}	5%
Utility function parameter, ε	0.5
Utility function parameter, β	0.5
Traffic period, Y	24hours
Time interval, T	1hour
SA's T_{max}	100
Number of channels, N_{ch}	100
Channel bandwidth, w	200KHz

5.5 Simulation Scenarios

The following five scenarios will be carried out to evaluate the performance of the proposed algorithms of the framework:

1. Radio Side PRRH Switch off
2. Baseband workload consolidation in the BS cloud
3. Combination of PRRH switch off algorithm (H-C-RAN) and baseband workload consolidation schemes of SA and GA.
4. Advanced CAC
5. Saving energy in the mobile device

This following sub sections describe these five scenarios in details.

5.5.1 Scenario 1: Radio Side PRRH Switch Off

These scenario involves PRRH switching off only at the radio side. The aim of this scenario is to show how the proposed RPRRH switch off scheme compares to others schemes in terms of the number of PRRHs it can switch off, blocking probability, execution time and the effects on SINR and interference. Here the results of the PRRH switch off algorithm are compiled where a baseline CAC mechanism is implemented. The proposed PRRH switch off scheme will be presented as *H-C-RAN* in the results section. The *H-C-RAN* scheme is compared with three other schemes that currently offer better performance; (i) the standard system called *LTE-A HetNet* which consists of 19 wraparound macro BSs overlaid by 12 pico BSs per macro BS each having its own BBU server. In *LTE-A HetNet*, there is no switching off of BSs or BBUs. (ii) a BS switch off algorithm called fixed progressive dynamic switching off (*F-PDSO*) proposed in [76]. In *F-PDSO*, only the PRRHs are

switched OFF and the BBUs are still decentralised to their respective coverage areas. In *F-PDSO*, there is no switching off of BBUs, (iii) a RAN as a Service H-CRAN based scheme proposed in [14] called *RANaaS* which switchs the PRRHs. The value of T here is set to 1 hour with the value of Y set to 24hours.

5.5.2 Scenario 2: Baseband Workload Consolidation in the BS Cloud

This involves the reduction of energy consumption only in the BS cloud by minimising the number of active BBU servers using the proposed SA based and GA based schemes. The motivations for proposing two schemes of SA and GA is to be able to compare them and select the most suitable one for our proposed architecture. The aim of this scenario is to show the effects of the proposed scheme on the number of active GPPs it can switch off, statistical multiplexing gain, GPP utilisation and execution time. The proposed schemes will be compared with the following schemes that offer better performance than other proposed schemes, baseline bin packing schemes (NF, FF, FFD and FBP), baseline LTE- A HetNet scheme where there is no BBU switch off, the RANaaS scheme where there is BBU switch off and the F-PDSO scheme where there is no BBU switch off.

5.5.3 Scenario 3: Combination of PRRH Switch off Algorithm (H-C-RAN) and Baseband Workload Consolidation Schemes of SA and SA.

After the performance of the proposed H-C-RAN (scenario 1) and the workload consolidation schemes of SA and GA (scenario 2) have been evaluated

separately, in this scenario they are combined for further enhancement of performance with the main aim to show the effects on overall network power consumption, energy efficiency, blocking probability, execution time, and throughput. The combination of H-C-RAN and SA scheme will be termed H-C-RAN SA and the combination of the H-C-RAN and GA schemes will be termed H-C-RAN GA. The H-C-RAN SA and H-C-RAN SA schemes will be compared with the baseline LTE-A HetNet, the F-PDSO and the RANaaS schemes.

5.5.4 Scenario 4: Advanced CAC in the BS cloud

The previous scenarios 1 to 3 were based on a baseline CAC scheme which drops the connections when a certain threshold is met which is 100 users since there are 100 PRBs within an MRRH and 8 users within the PRRH due to its small coverage area. However as explained previously, it is imperative that a more sophisticated CAC algorithm is used to make the most use of the benefits of the proposed framework. Based on the simulation results of scenario scenario 2 and 3 which will be explained in Section 5.6.2 and 5.6.3, it was observed that the H-C-RAN SA outperforms all the other schemes in terms of energy consumption and energy efficiency as such it is the most suitable one for our architecture and in this scenario, the proposed advanced CAC in the BS cloud is applied within the H-C-RAN SA scheme which is chosen for its best performance.

Three schemes are compared for performance evaluation;

- Baseline CAC scheme on the H-C-RAN SA scheme which have been used in scenarios 1 to 3,

- Proposed Fuzzy CAC without pre-emption within H-C-RAN SA scheme,
- Proposed Fuzzy CAC with pre-emption within H-C-RAN SA scheme.

5.5.5 Scenario 5: Saving Energy in the Mobile Device

In this scenario, the settings of saving energy in the mobile device will be presented. In this scenario, only a single MRRH overlaid by four PRRHs is considered without loss of generality with up to 100 users in consideration. The proposed EMCC scheme will be compared with two other schemes;

- The scheme where only local computation is performed in the mobile device.
- The EECO scheme in [52].

The application to be considered is the electroencephalogram (EEG) tractor game in [53] that involves augmented brain-computer interaction. In order to play the EEG tractor game, each player needs to wear the MINDO-4S wireless EEG headset sensor that is connected to the smart phone. The MINDO-4S headsets sensor detects the brain states and sends the brain state to the mobile device for processing. In our case, some processing is done in the MEC server in the BS cloud. The application display on the mobile device shows all players on a ring surrounding a target object. In order to win the game, the player with great brain concentration abilities detected by the MINDO-4S sensor will cause the target object to be pulled towards him. Figure 5-1, shows the application model of the EEG tractor game from [114] as a graph, $G = \{V, E\}$, where V is a set of vertices denoting the application task modules and E is the edges which denotes data dependencies. EEG module is the sensor that sends EEG signals to the client. The client is the mobile device, the display

is also in the mobile device. The concentration calculator determines the brain state of the user from the sensed EEG signal values and calculates the concentration level. This module informs the client module about the measured concentration level so that the game state of the player on the display can be updated. Coordinator works at the global level and coordinates the game between multiple players that may be present at geographically distributed locations. The Coordinator continuously sends the current state of the game to the Client module of all connected users. The simulation was performed using iFogSim simulation tool [114]. The iFogSim is a toolkit for modelling and simulation of resource management techniques in the IoT and MEC. Table 13 shows the settings used for the mobile application considered. The applications that were shared between the mobile device and the MEC server are the concentration calculator and the coordinator.

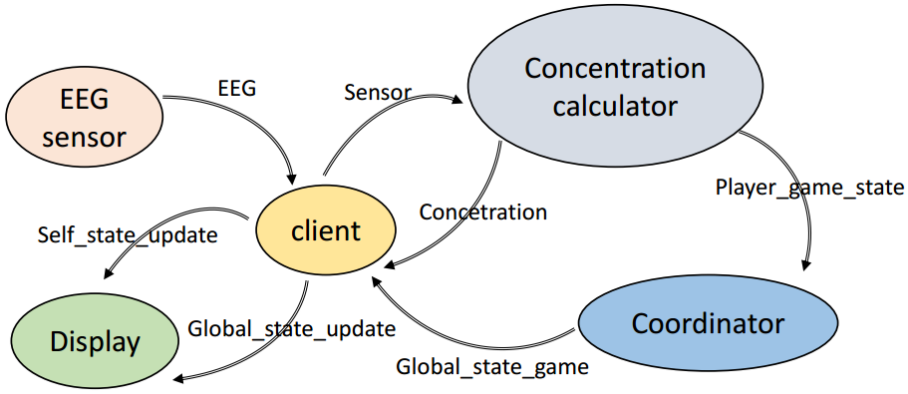

Figure 5-1. Application model for EEG game [114].

Table 13. Description of inter-module settings [114].

Task type	CPU length (MIPS)	Size (bytes)
EEG	2000	500
Sensor	3500	500

Player_game_state	1000	1000
Concentration	14	500
Global_state_game	1000	1000
Global_state_update	1000	500
Self_state_update	1000	500

5.6 Results Evaluation

The results of the proposed H-C-RAN framework and performance analysis are provided in this section. All the results of the five scenarios introduced in Section 5.5 will be presented in this section. The traffic profile used for the simulation scenarios is shown in Figure 5-2.

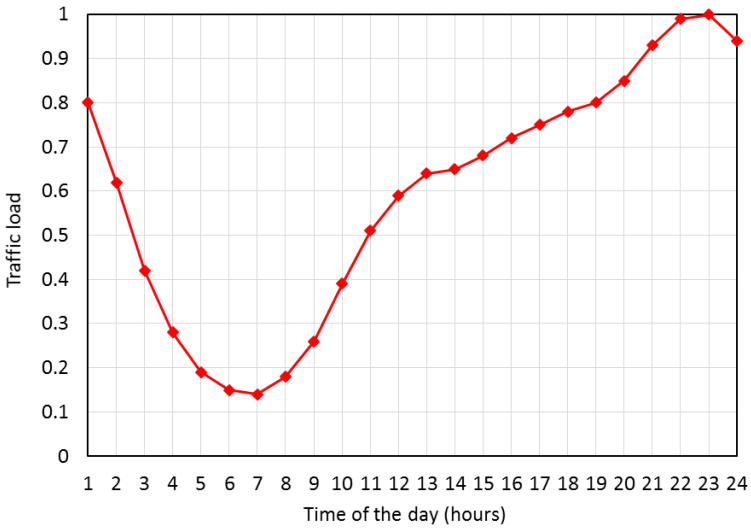

Figure 5-2. Traffic profile.

5.6.1 Results for Scenario 1: Radio Side PRRH Switch Off

5.6.1.1 Effects on the number of active PRRHs

Figure 5-3 shows the number of active PRRHs over 24 hour period. It shows that while for all the other schemes, the number of active PRRHs tracks the

traffic profile variation very well, the LTE-A HetNet scheme has a constant number of PRRHs since there is no switching off of the BSs. At low traffic periods, compared to the LTE-A HetNet scheme, the H-CRAN, RANaaS and F-PDSO has 5, 9 and 29 PRRHs active which correspond to 97%, 96% and 87% of the PRRHs switched off respectively. At peak traffic periods, compared to the LTE-A HetNet scheme, the H-CRAN, RANaaS and F-PDSO has 198, 225 and 228 active PRRHs which correspond to 13%, 1% and 0% of the PRRHs switched off respectively.

Figure 5-4 shows the number of PRRHs during low traffic, peak traffic and daily average. On a daily average, compared to the LTE-A HetNet, the H-CRAN, RANaaS and F-PDSO has 127, 152 and 164 active PRRHs corresponding to 44%, 33% and 28% respectively. The LTE-A HetNet scheme performs poorly with constant number of PRRHs (228=19*12 to be precise) since all PRRHs are always kept active irrespective of traffic variation. The proposed H-C-RAN scheme performs better than the rest in all the cases because the network state information (RRH load, interference) information is readily available in the centralised controller in the GCC making it easy for the PRRH switch off algorithm to be implemented whereas on the other schemes, the network state information is not easy to get as it is decentralised.

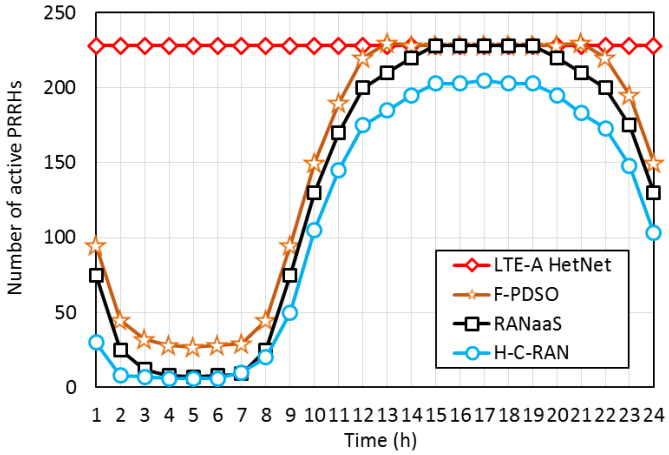

Figure 5-3. Number of active PRRHs over 24 hour period.

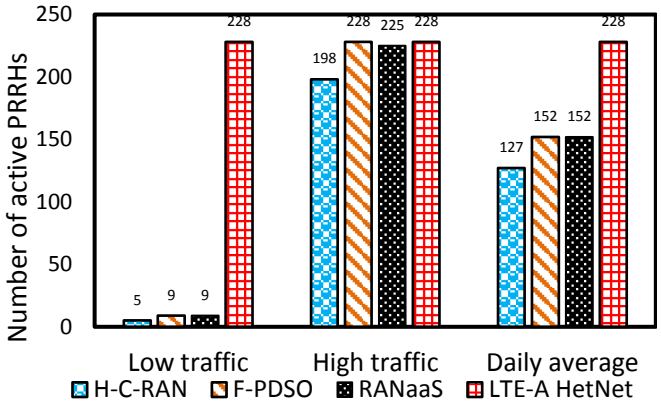

Figure 5-4. Number of active PRRHs during low and peak traffic and a daily average.

Figure 5-5 shows the effects of normalised traffic load on the number of active PRRHs for all the four schemes. It can be observed that for the LTE-A HetNet scheme, the number of PRRHs are a constant 228 since no PRRHs are switched off whereas for the other three schemes, the number of active PRRHs increases with the increase in traffic load as more users need to be served and as such more PRRHs need to be switched on. Therefore the

proposed H-C-RAN scheme is the best with an average of 45% better performance than the baseline LTE-A HetNet scheme.

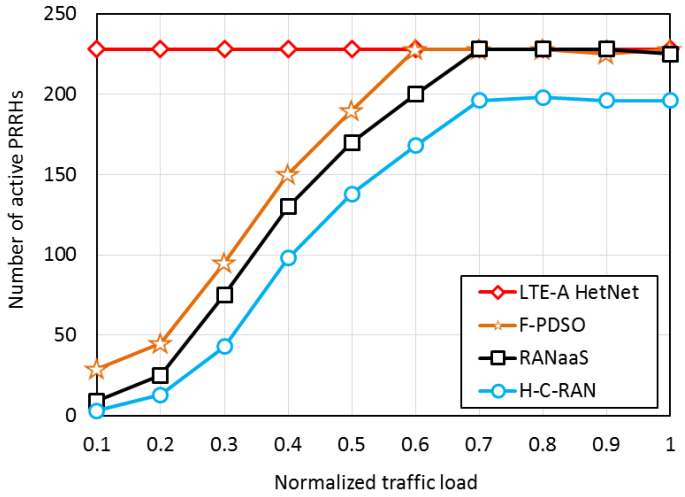

Figure 5-5. The effects of normalised traffic load on the number of active PRRHs.

5.6.1.2 Effects on the blocking probability.

Figure 5-6 shows the effect of traffic load on the average blocking probability. As the traffic load increases, the blocking probability also increases as more requests from users utilises more radio channels (PRBs).

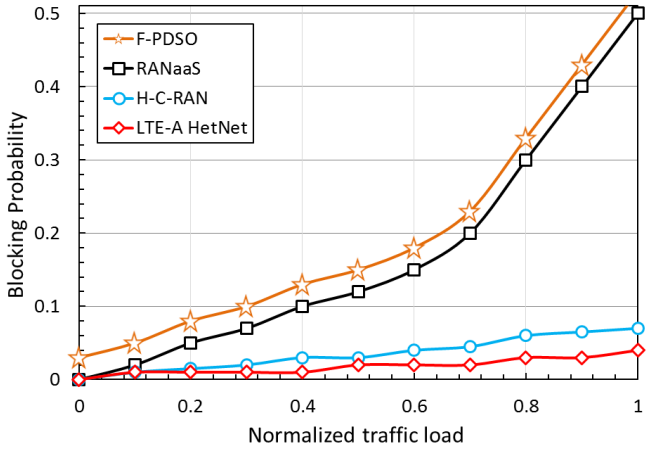

Figure 5-6. The effects of the number of users on the blocking probability.

The LTE-A HetNet scheme performs better than all the other schemes with an average blocking probability of 2% because there are no coverage holes and chances of signal outages are slim since there is no switching off of BSs. On average, the H-C-RAN, RANaaS and F-PDSO has blocking probability of 2.7%, 17% and 20% respectively. The H-C-RAN performs much better than the RANaaS and F-PDSO schemes with the average blocking probability of 2.7% within the acceptable QoS blocking probability limits of 5%. This is because in H-C-RAN, when the PRRHs are switched off, the nearby PRRHs or the underlying MRRH take over the users of the switched off PRRH through handover. The RANaaS and the F-PDSO performs poorly with a high average blocking probability of 17% and 20% respectively which are greater than the acceptable limit of 5%. The blocking probability for the RANaaS and F-PDSO scheme increases sharply due to the frequent switching off of BSs which create many coverage holes and signal outage for users.

Figure 5-7 shows the effects of increasing the number of PRRHs per MRRH on the blocking probability.

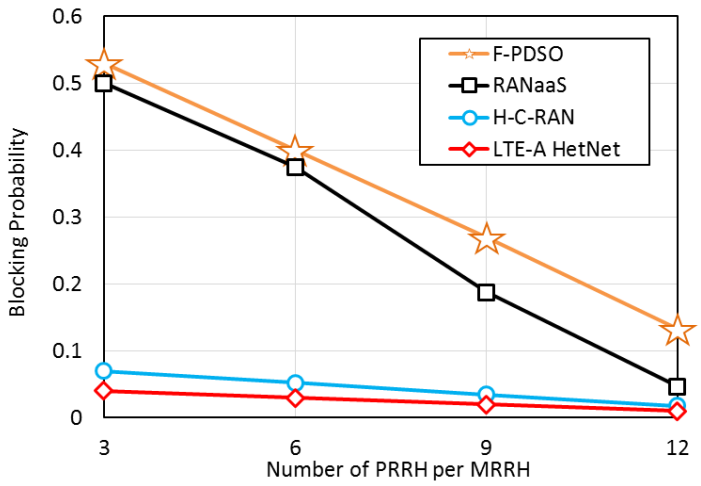

Figure 5-7. The number of active PRRHs per MRRH on blocking probability.

For all the schemes, as the number of PRRHs per MRRH increases, the blocking probability decreases because the coverage holes are decreased as more RRHs are available to serve the mobile users in the coverage area, as such, the outage is minimised.

For LTE-A HetNet and the H-C-RAN schemes, the blocking probability is an average of 2.5% and 4.4% respectively which is within the 5% blocking probability limit. The slope decreases slightly and almost constant with the LTE-A HetNet scheme because in the LTE-A HetNet scheme, there is no BS switching off as such, there are no chances of coverage holes. The LTE-A HetNet performs only 1.9% slightly better than the H-C-RAN scheme because in the latter, there is PRRH switching off which create chances for coverage holes that will result in increased blocking probability. Also, the PRRH switching off involves handing over users from the PRRH to be switched off to other RRHs and this increases chances of blocking probability. The F-PDSO and the RANaaS have an average blocking probability of 33% and 28% which is above the blocking probability limit of 5%, with a steep slope because both schemes involve frequent switching off of BSs which increases chances of blocking probability as users have to be handed over frequently from one BS to another.

5.6.1.3 Simulation times of the BS switch off schemes.

Figure 5-8 shows the execution times for all the schemes during the BS switching off. The execution time for the schemes seems to track/follow the traffic profile except for the LTE-A HetNet scheme which does not switch off any RRH/BS.

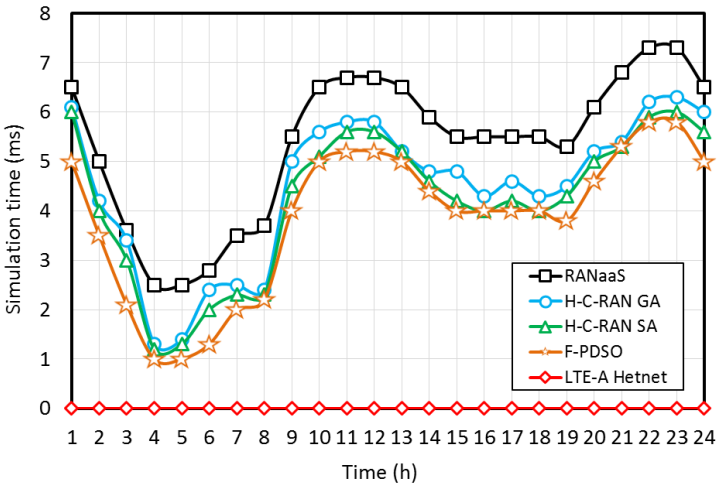

Figure 5-8. The simulation time for all the schemes.

The maximum simulation time for the RANaaS, H-C-RAN, F-PDSO and LTE-A HetNet schemes is 7.3ms, 6.3ms, 5.8ms and 0ms respectively. For the LTE-A HetNet scheme, the time is negligible as there is no switching off of BSs involved. The proposed H-C-RAN scheme performs better than the RANaaS scheme because in H-C-RAN, the state information is readily available in the centralised controller for making the switch off decision whereas in the RANaaS, the state information is decentralised in the BSs which makes the switching off scheme take time to collect the state information and make decisions. On the other hand, the H-C-RAN scheme is outperformed by the F-PDSO scheme because in F-PDSO, switches off fewer number of RRHs compared to the H-C-RAN scheme.

Figure 5-9 shows the effects of increasing the number of PRRHs per MRRHs on the simulation time on all the schemes and it can be observed that for all the other schemes except the LTE-A HetNet scheme, as the number of PRRHs per MRRH is increased, the execution time for the switch off scheme also increases because more PRRHs means the switch off algorithm has to

be executed with many PRRHs as inputs which will take time. On average, the LTE-A HetNet, F-PDSO, RANaaS and H-C-RAN have the average simulation time of 0ms, 3.6ms, 4ms and 3ms. The LTE-A HetNet scheme has a constant simulation time of 0ms since there is not switching off of RRHs. The H-C-RAN performs better (3ms) than the RANaaS scheme (4ms) because the state information required to switch off the PRRHs is readily available in the centralised controller.

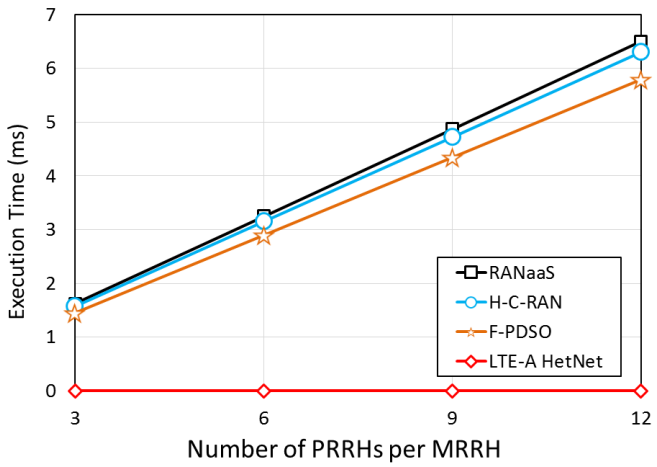

Figure 5-9. The number of PRRHs per MRRH versus maximum simulation time of the switch off algorithm.

5.6.1.4 Effects on user SINR and interference.

Figure 5-10 shows the effects of normalised cell traffic load on the SINR of a user at the edge of the cell and it can be observed that for all the schemes, as the traffic load within the cell increases, the SINR of the user decreases because of increased interference with the increase in the number of users within the cell. On average, the H-C-RAN, RANaaS, F-PDSO and the LTE-A HetNet schemes have an average SINR values of 1dB, -0.3dB, -1.5dB and -2dB. According to [115], the minimum SINR in LTE is from -6.7dB to 22.7dB.

Values below -6.7dB will lead to signal outage and poor communication channel.

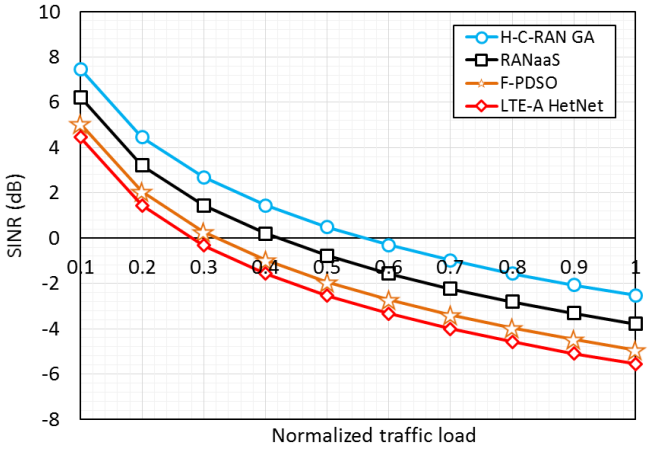

Figure 5-10. The effects of normalised cell traffic load on SINR of a user at the edge of the cell.

The proposed H-C-RAN scheme performs better than all the other schemes with an maximum SINR of 7dB which is within the SINR limit because H-C-RAN switches off as many PRRHs as possible as such the interference from PRRHs is decreased hence improving the SINR of a user. The LTE-A HetNet scheme has lowest average SINR (4dB) compared to all the other schemes because there is no BS switching off as such there is more interference.

5.6.2 Results for Scenario 2: Baseband Workload Consolidation in the BS cloud.

5.6.2.1 Effects on the number of active GPPs/BBUs.

Figure 5-11 shows the total number of active GPPs that are kept active in the BS cloud against the normalised traffic load. Nine schemes were compared to find out how they perform in terms of GPP reduction as the traffic load increases. It can be seen from the figure that for the LTE-A HetNet and F-

PDSO schemes, the number of active GPPs is a constant value of 247 (228 PRRHs plus 19 MRRHs) since there is no switching off of GPPs in these two schemes as each cell has its own individual GPP for baseband processing as such, the number of BBUs equal the number of RRHs and the BBUs are always kept active irrespective of traffic load from the radio side.

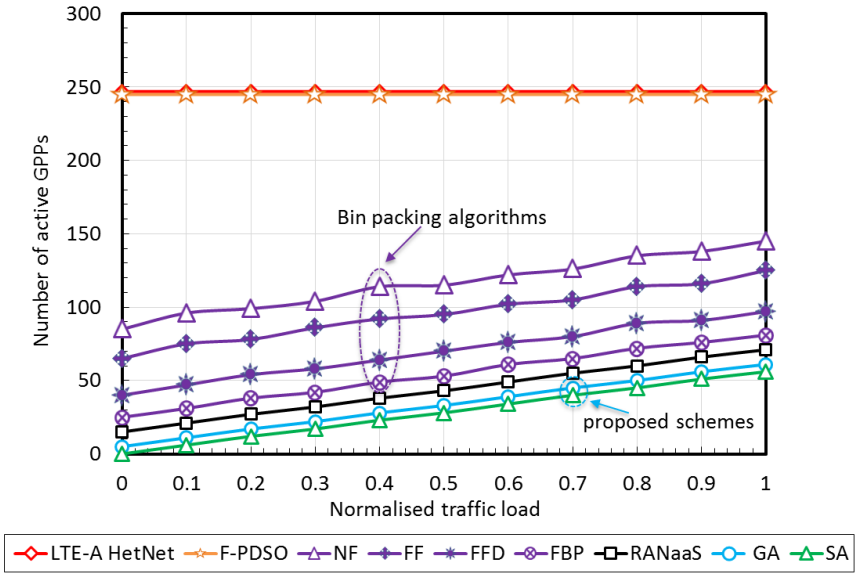

Figure 5-11. Total number of active GPPs in the BS cloud vs normalised traffic load.

For other schemes, as the traffic load increases, the number of active GPPs required also increases due to the fact that there is GPP switching on and as such the traffic load increases, more baseband processing power is required which cause more GPPs to be activated. Figure 5-12 shows the number of active GPPS for low traffic, peak traffic and daily average. During low traffic periods, the SA, GA, RANaaS, FBP, FFD, FF and NF switches off 241, 236, 226, 222, 207, 182 and 162 GPPs respectively which compared to the baseline LTE-A scheme, correspond to 98%, 96%, 91%, 90%, 84%, 74% and 66% performance improvement respectively. During peak traffic periods, compared to the baseline LTE-A HetNet scheme, the SA, GA, RANaaS, FBP,

FFD, FF and NF outperforms by 77%, 75%, 71%, 67%, 61%, 49% and 41% respectively and on daily average, 84%, 83%, 81%, 78%, 72%, 62% and 53% respectively. In all the traffic load cases, the proposed SA and GA outperforms all the other schemes because more GPPs are switched off in these schemes. On average, the SA scheme performs better than the GA scheme by 1% as the SA scheme switches off more GPPs since the SA accepts even worst case scenarios to avoid being stuck at local optimum.

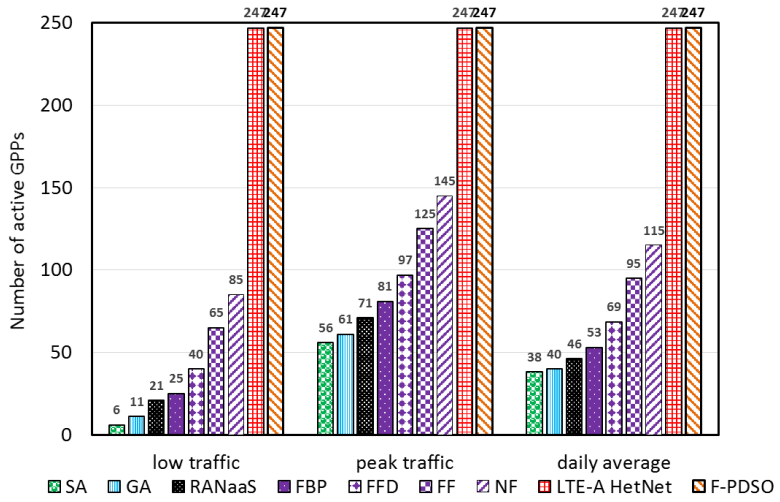

Figure 5-12. Number of active GPPs during low traffic, peak traffic and daily average.

5.6.2.2 Effects on the statistical multiplexing gain.

Figure 5-13 shows the effects of increased traffic load on the statistical multiplexing gain (θ) with the baseline LTE-A HetNet as a benchmark with a constant number of GPPs (247).

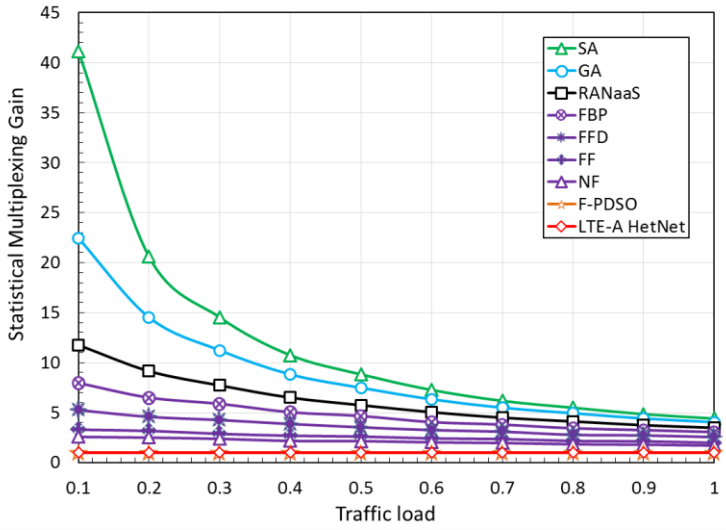

Figure 5-13. Statistical multiplexing gain versus normalised traffic load.

When θ for a particular scheme is greater than one, it means that the scheme performs better than the baseline LTE-A HetNet scheme due to switching off some GPPs. For the F-PDSO scheme, θ is one, as F-PDSO has the same number of active GPPs as the LTE-A HetNet scheme since there is no GPP switching off in that scheme. The proposed schemes SA and GA have high θ of 41 and 22 respectively during low traffic periods because during this periods, more GPPs are switched off with the SA scheme switching off 5 GPPs more than the GA and, as the traffic load increases, θ decreases because more GPPs are switched on making the denominator of θ larger. On average, θ for SA, GA, RANaaS, FBP, FFD, FF, NF and F-PDSO is 12, 9, 6, 5, 4, 3, 2 and 1. On average, the SA and GA performs better than all the other schemes since more GPPs are switched off.

5.6.2.3 Effects on the GPP utilisation.

Figure 5-14 shows the average GPP utilisation versus normalised traffic load in the BS cloud. It can be observed that for all the schemes, as the traffic load

increases, the average GPP utilisation also increases because more baseband traffic is being processed in the GPPs. On average, the LTE-A HetNet, F-PDSO, RANaaS, GA and SA has an average GPP utilisation of 34%, 34%, 75%, 80% and 83% respectively. The proposed SA perfoms better than the other schemes which has high utilisation because in SA, fewer servers are processing all the traffic from all cells in C-RAN since GPPs are shared in the BS cloud. The SA scheme switches off more GPPs than the GA scheme.

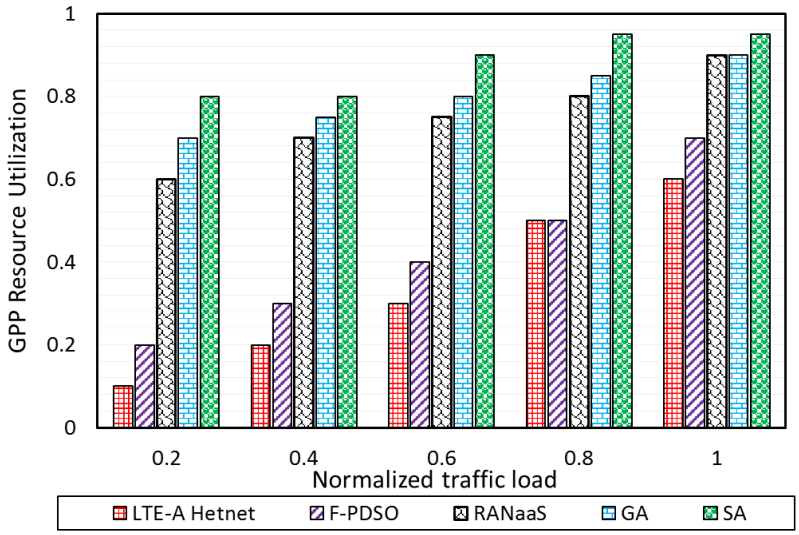

Figure 5-14. The average GPP utilisation versus normalised traffic load in the BS cloud.

The LTE-A HetNet and the F-PDSO has low GPP utilisation for the same traffic load because the GPPs/BBUs are decentralised and not shared, so that there is no consolidation of BBUs, which causes them to be underutilised.

5.6.2.4 Simulation times of the GPP switch off schemes.

Figure 5-15 shows the simulation time versus time of the day for all the schemes. For the F-PDSO and LTE-A HetNet schemes, the simulation time is zero because there is no switching off of GPPs/BBUs. For the RANaaS, GA and SA, the maximum simulation time for the switching off the GPPs is 8ms, 7.2ms and 6.8ms and with a daily average of 6ms, 5.2ms and 4.7ms. The proposed schemes of GA and SA perfoms better than the RANaaS scheme because the proposed schemes consist of a centralised controller (GCC) which make decisions on which GPP to switch off depending on the readily available global state information (e.g. GPP CPU utilisation) of each GPP whereas the RANaaS is dependent on distributed controllers which does not have access to the global state information in the BS cloud.

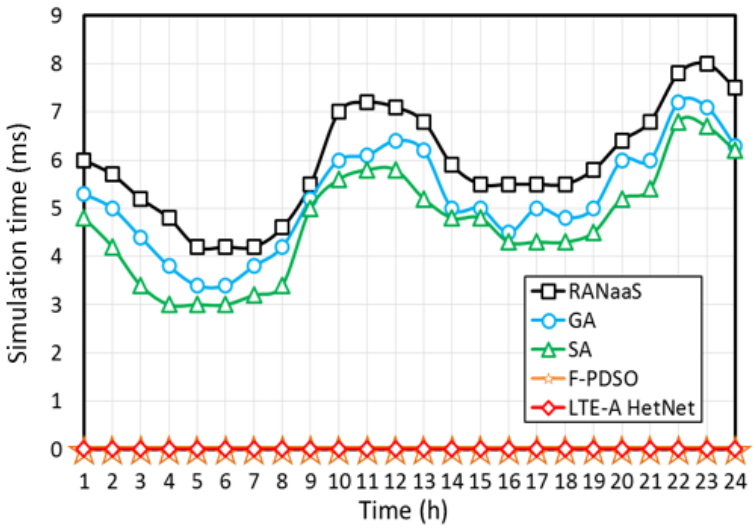

Figure 5-15. Simulation time versus time of the day for all the schemes.

5.6.3 Results for Scenario 3: Combination of PRRH Switch Off Algorithm (H-C-RAN) and Baseband Workload Consolidation Schemes of SA and SA

5.6.3.1 Effects on total network power consumption.

Figure 5-16 shows the network power consumption versus normalised traffic load in the network while Figure 5-17 summarises by showing the network power consumption during low traffic, high traffic and daily average. It can be seen that for all the schemes, as the network traffic load increases, the total network power consumption also increases because more traffic utilises more resources both in the radio side and in the BS cloud which results in increased power consumption. During low traffic periods, the LTE-A HetNet, F-PDSO, RANaaS, H-C-RAN GA and H-C-RAN SA has power consumptions of 20KW, 8KW, 5.5KW, 4.2KW and 3.6KW respectively whereas during high traffic periods, the power consumption is 40KW, 35KW, 32KW, 30KW and 29KW respectively, and on daily average, the power consumption is 30KW, 21KW, 18KW, 16.5KW and 15.6KW respectively.

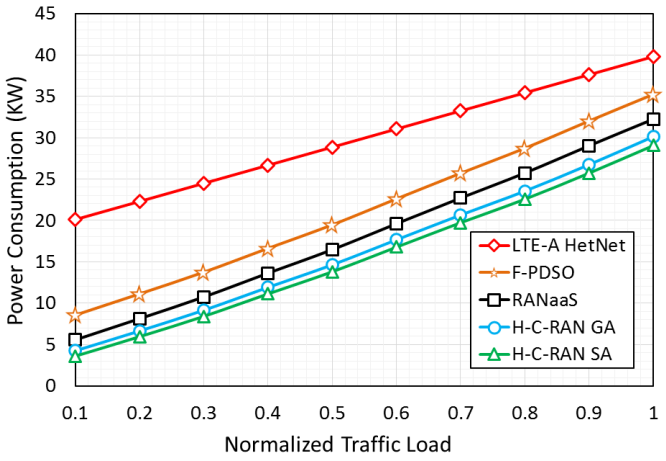

Figure 5-16. Total network power consumption versus normalised traffic load in the network.

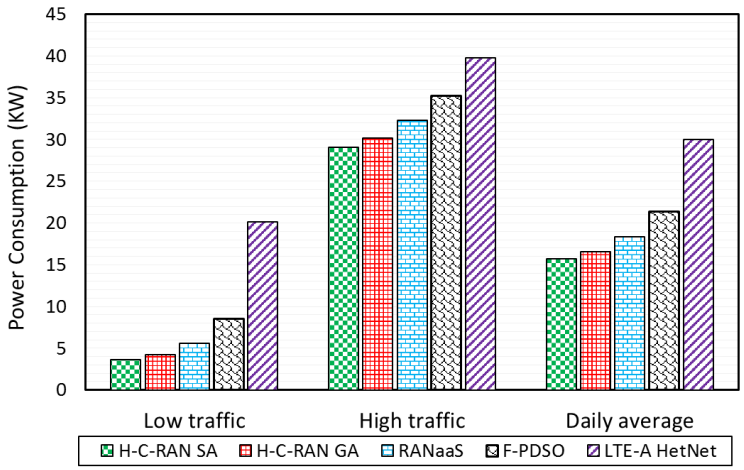

Figure 5-17. Network power consumption during low traffic, high traffic and daily average.

Compared to the LTE-A HetNet, F-PDSO and RANaaS, the H-C-RAN GA performs better by 45%, 21% and 8% whereas the H-C-RAN SA performs better by 48%, 26% and 14%. The LTE-A HetNet consumes more energy than all the other schemes because all the RRHs and BBUs are kept active all the time irrespective of traffic load. The F-PDSO consumes more power compared to the proposed schemes because there is only RRH switching off and no switching off of the BBU servers. The RANaaS consumes more power compared to the proposed schemes since in RANaaS, even though there is RRH switching off and BBU switching off, less servers are switched off compared to the proposed schemes. For the proposed schemes, less power is consumed in H-CRAN SA (3% less) than in H-CRAN GA because more BBUs are switched off in H-CRAN SA.

Figure 5-18 shows the total power consumption in the network during 24 hours period. It can be seen that for all the schemes, the power consumption tracks the traffic profile very well such that during low traffic periods in the day, less

energy is consumed whereas during peak periods, more equipment are activated, hence consuming a significant amount of energy.

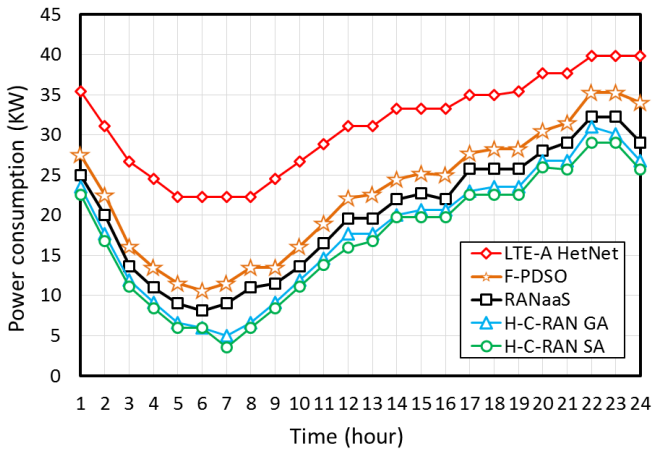

Figure 5-18. Power consumption in the network during 24 hours.

Figure 5-19 shows the effect of increasing the number of PRRHs per MRRH on the network power consumption. As the number of PRRHs per MRRH increases, the power consumption in the network also increases since more PRRHs consumes more energy and also they results in the creation of more vBBUs which results in more GPPs/BBUs to be turned on in the BS cloud.

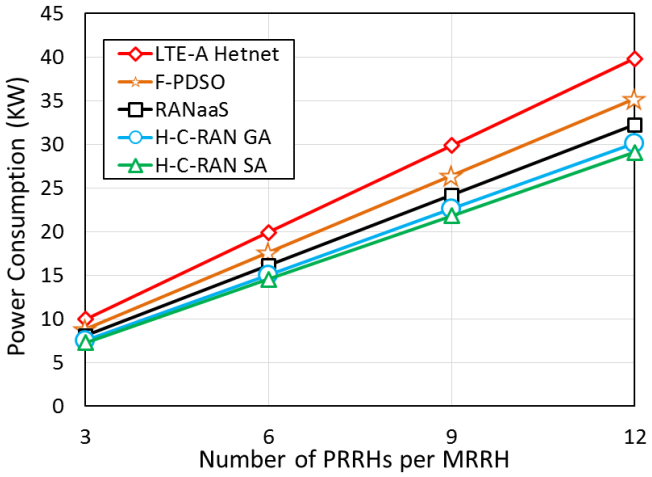

Figure 5-19. The effects of increasing the number of PRRHs per MRRH on the network power consumption.

Figure 5-20 shows the area power consumption in the network. As observed, when the traffic load in the network increases, the area power consumption for all the schemes also increases because the power consumption per km^2 also increases with the increase in traffic load.

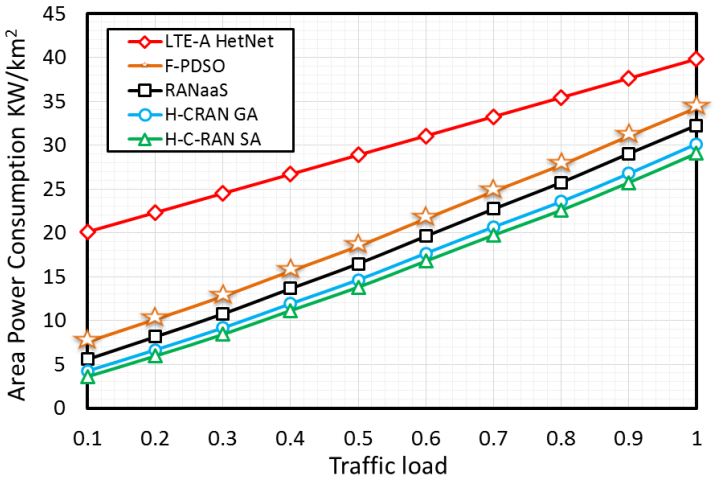

Figure 5-20. Area power consumption in the network.

On average, the LTE-A HetNet, F-PDSO, RANaaS, H-C-RAN GA and H-C-RAN SA has area power consumption of 30KW/km^2, 20KW/km^2, 18KW/km^2, 17KW/km^2 and 16KW/km^2. The LTE-A HetNet has high area power consumption than all the other schemes because it consumes a lot of power per km^2 since all the BSs are kept active irrespective of traffic load. Compared to the LTE-A HetNet, F-PDSO and RANaaS, the proposed schemes of H-C-RAN GA have minimised area power consumption by 43%, 15% and 5.5% and the proposed H-C-RAN SA by 47%, 20% and 11%.

Figure 5-21 shows the effect of moving a certain percentage of baseband processing to the cloud (by varying $\beta_{BB} \in [10\%, 50\%, 100\%]$ on the overall power consumption for H-CRAN SA scheme (chosen for best performance). The advantage of leaving some baseband tasks in the radio side on RRHs is to reduce the high bandwidth baseband signals transported between the BS

cloud and the radio side that can cause high cost in fibre. As shown in the graph, when 10%, 50% and 100% baseband is moved to the BS cloud, power savings of 41%, 42% and 48% are achieved respectively since more workload is consolidated and shared between GPPs and idle GPPs are turned off.

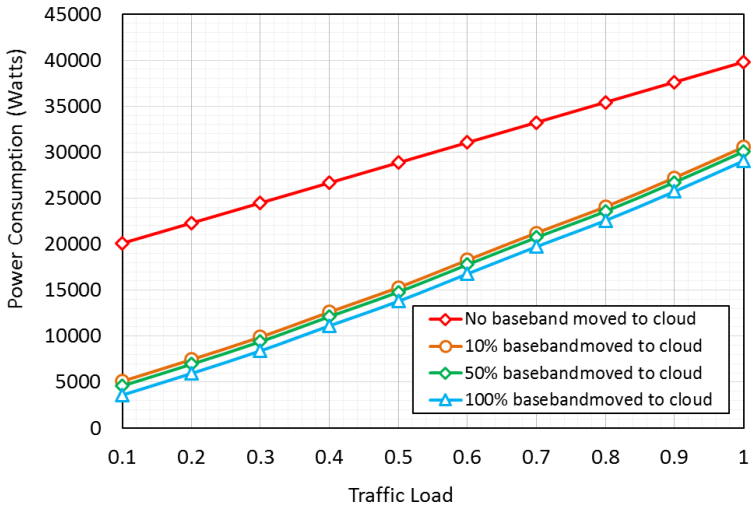

Figure 5-21. The effect on power consumption of moving some percentage of baseband to the BS cloud for the H-CRAN SA scheme

5.6.3.2 Effects on total network throughput.

Figure 5-22 shows the effects of increased traffic load on the total network throughput whereas Figure 5-23 summarises these results showing the network throughput during low traffic, high traffic and daily average.

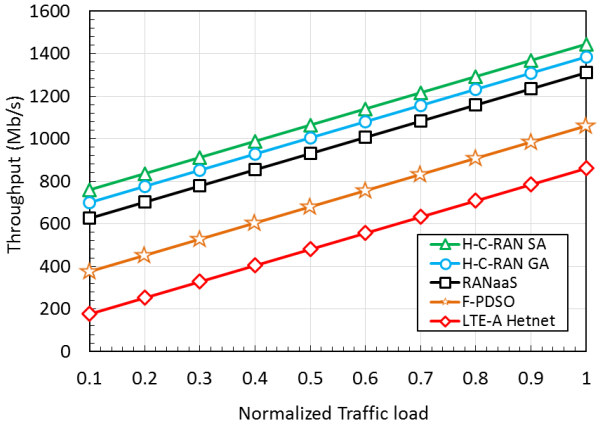

Figure 5-22. Effects of normalised traffic load on the total network throughput.

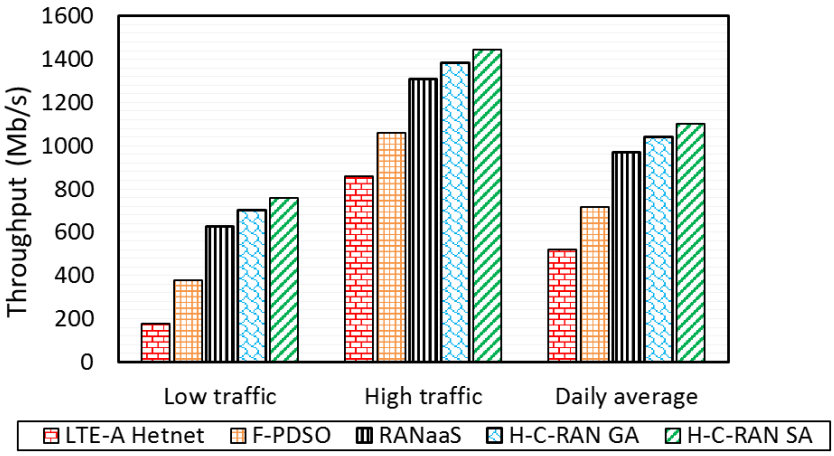

Figure 5-23. Network throughput during low traffic, high traffic and daily average.

It can be observed that when the traffic load increases, the throughput for all the schemes also increases because more users are transmitting data in the BS cloud hence increased the throughput. During low traffic periods, the H-C-RAN SA, H-C-RAN GA, RANaaS, F-PDSO and LTE-A HetNet has throughput of 760Mb/s, 700Mb/s, 626Mb/s, 376Mb/s and 176Mb/s respectively whereas during peak traffic, the throughput is 1444Mb/s,1384Mb/s,1310Mb/s, 1060Mb/s and 860Mb/s respectively and on average, the throughput is 1102Mb/s, 1042Mb/s,968Mb/s, 718Mb/s and 518Mb/s respectively. Compared to the other schemes, the LTE-A HetNet scheme performs poorly with low throughput due to increased interference in the system as no BSs are switched off to reduce interference since interference limit throughput. Compared to the RANaaS and F-PDSO scheme, the proposed H-C-RAN SA performs better by 14% and 53% respectively, whereas the and H-C-RAN GA perfoms better by 8% and 45% respectively because the proposed schemes switches off more PRRHs as such the interference that limit throughput is reduced. The H-C-RAN SA perfoms better than the H-C-RAN GA by 6%

because, H-C-RAN SA consolidate the user traffic into fewer number of GPPs as such this improves utilisations and increases utilisation.

Figure 5-24 shows the effects of the increase in the number of PRRHs per MRRH on the network throughput.

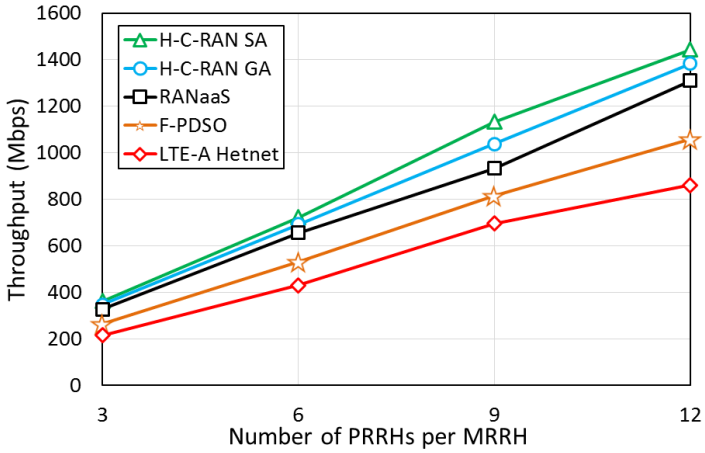

Figure 5-24. The effects of the increase in the number of PRRHs per MRRH on the network throughput.

It can be shown in the figure that as the number of PRRHs per MRRH increases, the throughput also increases for all the schemes because, more PRRHs increases the user datarate as PRRHs have high throughput compared to MRRHs.

5.6.3.3 Effects on energy efficiency.

Figure 5-25 shows the energy efficiency performance for increasing cell load. Compared to the LTE-A HetNet scheme, the H-C-RAN SA, H-C-RAN GA, RANaaS and F-PDSO perfoms better by a factor of 10, 9, 6 and 2, respectively during low traffic periods and by a factor of 2.5, 1.7, 1.5 and 1.7 respectively during peak periods.

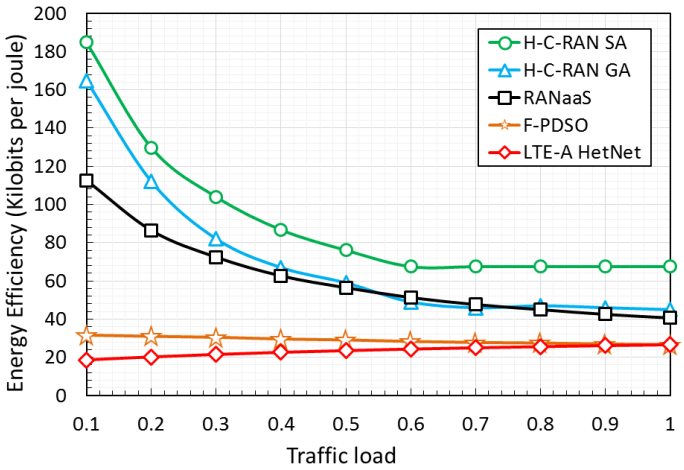

Figure 5-25. Network energy efficiency.

The energy efficiency significantly improved in the proposed H-CRAN schemes especially at low traffic periods because less power is consumed through workload consolidation as traffic from various cells is aggregated into fewer number of servers and as such making the denominator of the energy efficiency smaller. Also, the proposed H-C-RAN schemes have high throughput. The F-PDSO, RANaaS and LTE-A HetNet schemes has lower energy efficiency because they have lower throughput and consumes more power than the proposed schemes.

5.6.3.4 Effects on the blocking probability.

Figure 5-26 shows the blocking probability of all the schemes. It can be observed that during low traffic periods, the LTE-A HetNet, H-C-RAN GA, H-C-RAN SA, RANaaS and F-PDSO schemes has the blocking probability of 1%, 1%, 1%, 1% and 3% respectively and during high traffic periods the blocking probability is 2%, 4%, 5%, 48% and 50% respectively whereas on daily average, the blocking probability is 1%, 3.8%, 4.2%, 17% and 20% respectively.

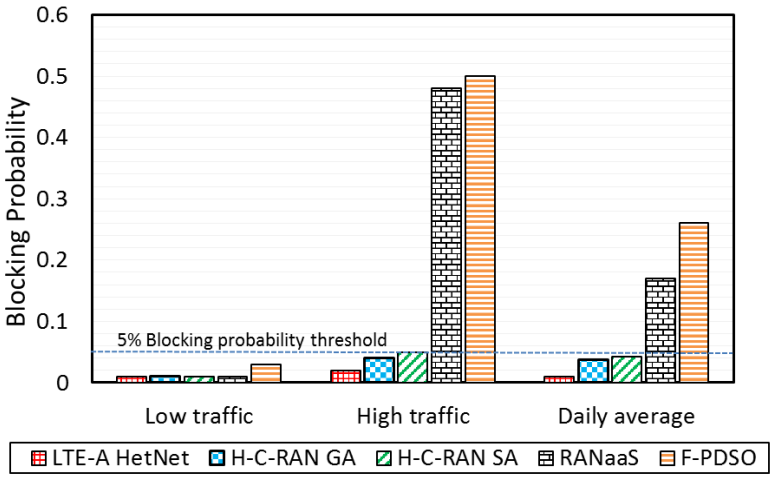

Figure 5-26. Blocking probability for all the schemes.

It can be observed that for both low traffic, high traffic and daily average, the proposed H-C-RAN SA and H-C-RAN GA schemes perfoms better than all the other schemes with blocking probability within the threshold of 5% because H-C-RAN SA and H-C-RAN GA are implemented in the centralised controller in the GCC that have access to the entire system state information (GPP utilisation, RRH utilisation, system interference) hence decision making of which PRRH and GPP to switch off is precise. The RANaaS and the F-PDSO has high blocking probability because the RANaaS scheme does not consider the trade-off between saving energy and QoS as such blocking probability was neglected, whereas the F-PDSO scheme involves frequent switch off of BSs with the state information decentralised and limited.

5.6.3.5 Comparison of Simulation times

Figure 5-27 shows the simulation time of all the schemes versus time. For the RANaaS, H-C-RAN GA, H-C-RAN SA, F-PDSO and LTE-A HetNet schemes, the maximum simulation time is 8ms, 7.2ms, 6.8ms, 5.8ms and 0ms respectively.

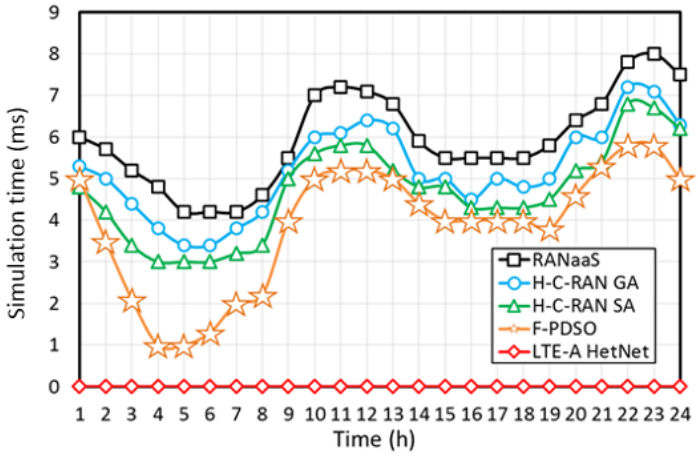

Figure 5-27. Simulation time of all the schemes versus time.

It can be observed that even though the proposed H-C-RAN GA and H-C-RAN SA consist of a combination of the H-C-RAN switch off scheme together with the GA and SA scheme, the simulation times for H-C-RAN GA and H-C-RAN SA follows the simulation time for the GA and SA schemes respectively. This is because the RRH switch off scheme (H-C-RAN) is executed at the same time as the GPP switch off scheme (SA or GA) as such the time that will be registered will be for the scheme with longest simulation time (SA or GA).

5.6.4 Results for Scenario 4: Advanced CAC in the BS cloud within H-C-RAN SA scheme.

5.6.4.1 Effects on the blocking probability

Figure 5-28 shows the blocking probability in the network for the baseline and proposed Fuzzy (with and without Pre-emption) CAC schemes within H-C-RAN SA scheme. Figure 5-29 shows the average blocking probability.

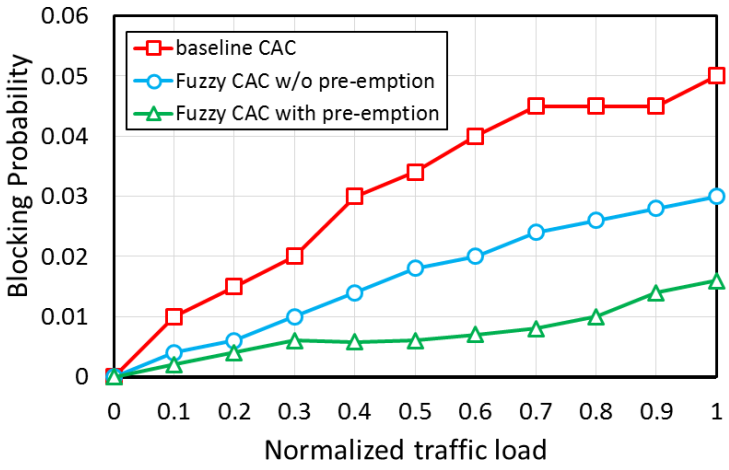

Figure 5-28. Blocking probability in the network within the H-C-RAN SA scheme.

It can be seen that for all the CAC schemes, as the traffic load increases, the blocking probability also increases because more requests occupy more resources in both the radio and the BS cloud. On average the blocking probability for baseline CAC, Fuzzy CAC without pre-emption and Fuzzy CAC with pre-emption schemes is 4.2%, 1.6% and 0.7% respectively.

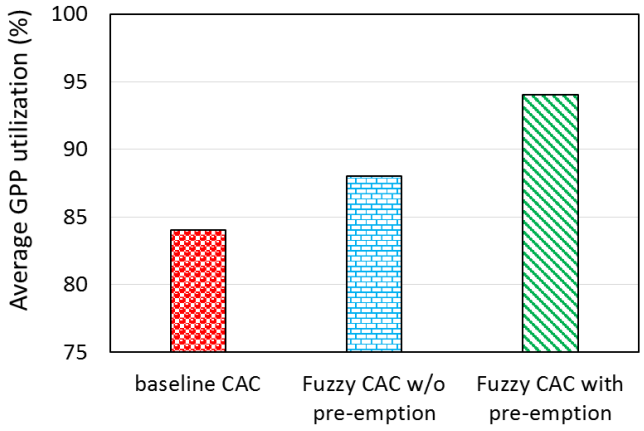

Figure 5-29. Average blocking probability for the CAC schemes in the network with H-C-RAN SA scheme.

The Fuzzy CAC without pre-emption scheme perfoms better than the baseline CAC scheme by 2.6% because fuzzy-logic avoids imprecisions and uncertainties found in the simple CAC scheme when performing CAC in the

BS cloud. The Fuzzy CAC with pre-emption scheme perfoms better than all the other schemes because, on top of avoiding imprecisions and uncertainties in the decision making, during the arrival of RT connections into the congested BS cloud, instead of the RT connection being blocked, the NRT connections are pre-empted from the operator RAN to the public RAN and also, during the arrival of NRT connections in the congested BS cloud, instead of the NRT connections being blocked, they are forwarded to the public cloud and this lowers the blocking probability significantly.

5.6.4.2 Effects on the GPP utilisation in the BS cloud

Figure 5-30 shows the effects of the CAC schemes on GPP utilisation within the BS cloud. It can be seen that with the introduction of the Fuzzy CAC without pre-emption, the average GPP utilisation increased by 4% from 84% to 88% when compared to the baseline CAC algorithm. This is because the Fuzzy CAC without pre-emption scheme has lower blocking probability than baseline CAC scheme resulting in more request being accepted in the BS cloud which will in turn result in increased utilisation.

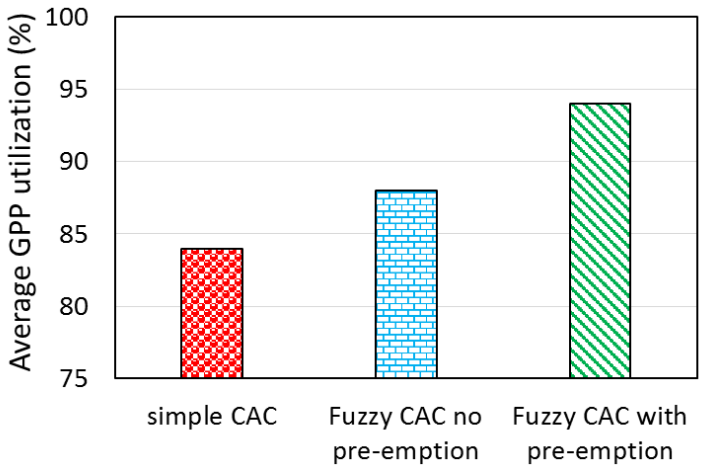

Figure 5-30. Effects of the CAC schemes on the BS cloud GPP utilisation.

The further introduction of pre-emption technique increases the GPP utilisation by 10% from 84% to 94% because pre-emption of connections to the public cloud means the blocking probability is decreased even more, which results in more GPPs processing more requests.

5.6.4.3 Effects on the H-C-RAN SA simulation time

Figure 5-31 shows the effects of the CAC schemes on the simulation time of the H-C-RAN SA scheme. The maximum response time for baseline CAC, Fuzzy CAC without pre-emption and Fuzzy CAC with pre-emption schemes is 6.8ms, 7ms and 7.2ms respectively. The introduction of fuzzy in Fuzzy CAC without pre-emption increased the response time of the H-C-RAN SA scheme by 2ms because more time is taken to perform the fuzzy rule base and the defuzzification within the fuzzy controller. The introduction of pre-emption in the proposed Fuzzy CAC with pre-emption scheme increased the simulation time of the H-C-RAN SA even more by 4ms because pre-emption involves removing the ongoing NRT connections from the BS cloud and outsourcing to the public cloud.

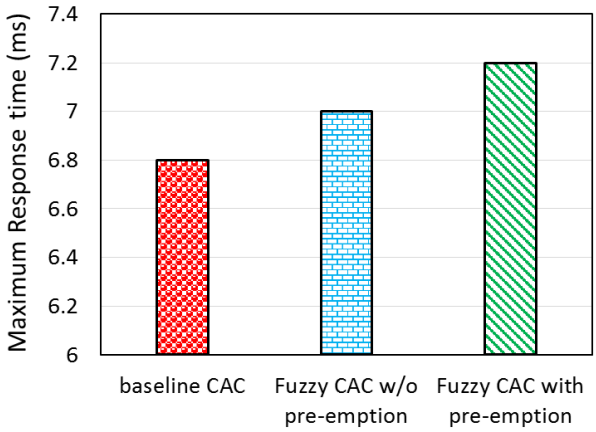

Figure 5-31. The effects of the CAC schemes on the simulation time of the H-C-RAN scheme.

5.6.4.4 Effects on the operator revenue

The operator revenue was calculated by using the blocking probability, for example, if the average blocking probability is 5% the revenue will be 95%. Figure 5-32 shows the operator revenue for all the schemes. The operator revenue for the baseline CAC scheme, Fuzzy CAC without pre-emption and Fuzzy CAC with pre-emption is 95.8%, 98.4% and 99.3% respectively.

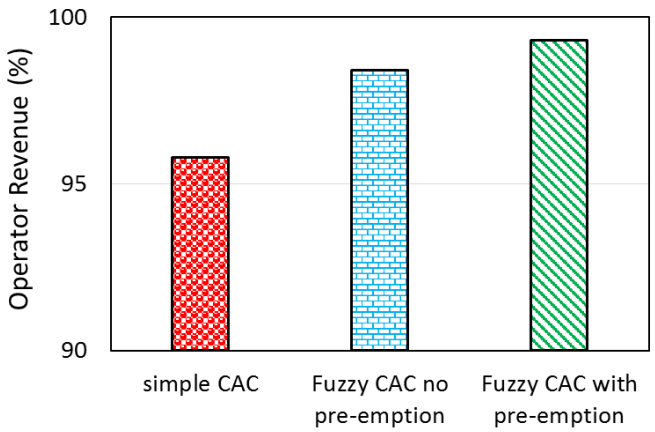

Figure 5-32. Operator revenue for all the schemes.

The revenue increased by 2.6% with the introduction of fuzzy logic in Fuzzy CAC without pre-emption because of low blocking probability in Fuzzy CAC without pre-emption scheme. The operator revenue increased even more by 3.5% with the implementation of pre-emption on top of fuzzy logic in Fuzzy CAC with pre-emption because more requests are accepted in both the public and the private cloud and this lowers blocking probability significantly resulting in increased revenue for the operator.

5.6.4.5 Effects on the network throughput

Figure 5-33 shows the effects of the CAC schemes on the network throughput while Figure 5-34 summarises Figure 5-33 by showing the average network

throughput. As shown in the diagram, the network throughput increases with the increase in traffic load since more traffic means more PRBs used which increases throughput. On average, the baseline CAC, Fuzzy CAC without pre-emption and Fuzzy CAC with pre-emption schemes have network throughput of 1102Mbps, 1157Mbps and 1432.6Mbps respectively.

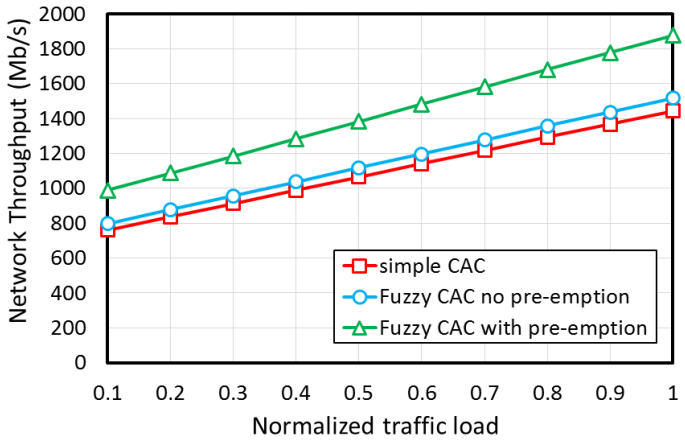

Figure 5-33. The effects of the CAC schemes on the network throughput.

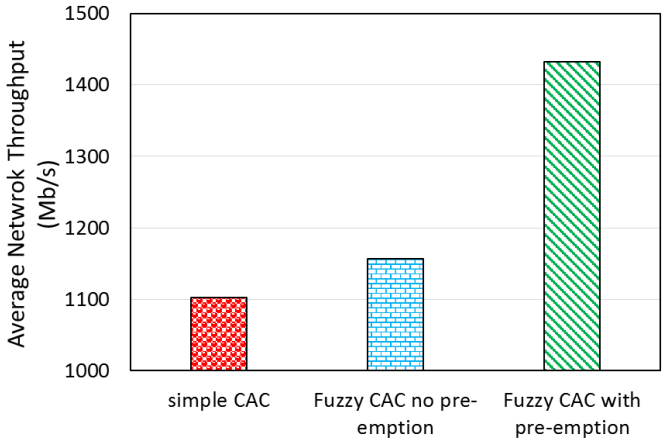

Figure 5-34. Average network throughput.

The introduction of fuzzy logic in Fuzzy CAC without pre-emption scheme caused an increase in the in throughput of 5% from 1102Mbps to 1157Mbps. This is because compared to the baseline CAC scheme, the Fuzzy CAC

without pre-emption scheme has high call acceptance rate as such more PRBs resources are allocated in the network which will cause an increase in throughput. The introduction of pre-emption together with fuzzy logic in Fuzzy CAC with pre-emption caused an increase of throughput by 30% from 1102Mbps to 1432.6Mbps because compared to the baseline CAC scheme with average blocking probability of 4.2%, the Fuzzy CAC with pre-emption scheme has an average blocking probability of 0.7% which results in high call acceptance rate which requires more PRBs which will increase the total network throughput.

5.6.4.6 Effects of the CAC schemes on the network power consumption

Figure 5-35 shows the total power consumption of the H-C-RAN SA scheme for all the CAC schemes with the increase in the traffic load. On average, the baseline CAC, Fuzzy CAC without pre-emption and Fuzzy CAC with pre-emption schemes cause the total power consumption of 15.7KW, 18KW and 19KW respectively. The Fuzzy CAC without pre-emption consumes 15% more power consumption compared to the baseline CAC scheme because the Fuzzy CAC without pre-emption accept more connections than the baseline CAC scheme and more connections consumes more power. The Fuzzy CAC with pre-emption consumes 22% more power compared to the baseline CAC because the Fuzzy CAC with pre-emption scheme accepts more connections both in the private and the public cloud which will increase the energy consumption in the network.

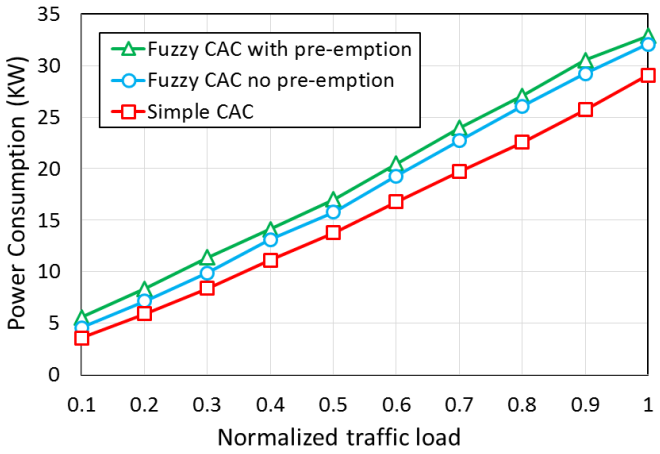

Figure 5-35. Total power consumption of the H-C-RAN SA scheme for all the CAC schemes.

5.6.5 Results for Scenario 5: Saving Energy in the Mobile Device.

5.6.5.1 Effects on total time of execution of the application task

Figure 5-36 shows the average time of execution on the Local computation, EECO and EMCC scheme. On average, the Local computation, EECO and EMCC schemes have execution times of 9.5ms, 4ms and 3.5ms respectively.

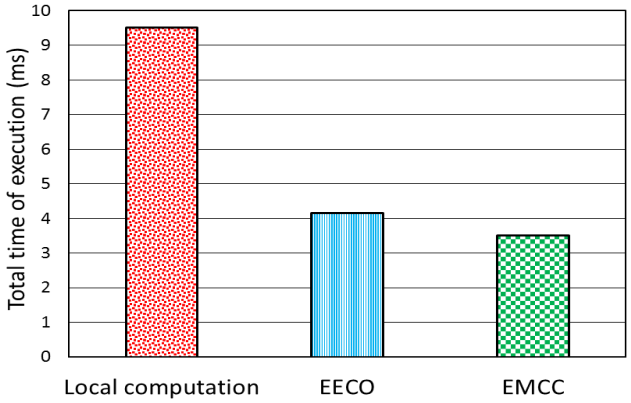

Figure 5-36. Average time of execution of all the schemes.

The EMCC scheme perfoms better than the Local computation scheme by 63% because even though the task data does not need to be transmitted

anywhere during local computation, the mobile device itself has very slow computation capability compared to the MEC computation capability which will results in more time being taken for execution of a particular task locally within the device. The EMCC scheme perfoms better than the EECO scheme by 12.5% because, in EMCC, user association to PRRHs is done by the criterion of CRE which cause more users to be associated to the PRRH which has high transmission datarate whereas the EECO schemes use maximum RSRP for user association which lead to less users associated and transmitting via the MRRH. Also, in EECO, tasks are executed in the MEC server but not in parallel whereas in EMCC, tasks are executed in parallel by the VMs in the MEC servers and this will reduce the time of execution.

Figure 5-37 shows the effects of increasing the MEC server speed on execution time of the proposed EMCC on various mobile devices. The figure shows that the execution time decreases as the MEC processing power increases since tasks are executed within a short period of time.

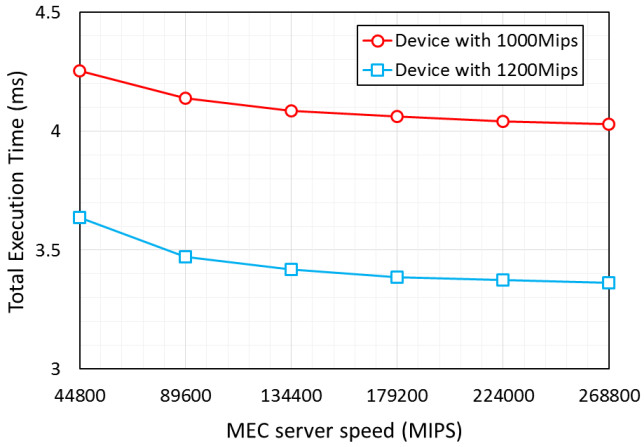

Figure 5-37. The effects of increasing the MEC server speed on execution time of the EMCC scheme on various mobile devices.

Also, the greater the mobile device's processing power, the less the execution time (the 1200Mips device improved execution time by 12% compared to the 1000Mips device) since more instruction are processed in the local device.

Figure 5-38 shows the effects of varying the task input data size on the execution time of the EMCC scheme on various mobile devices.

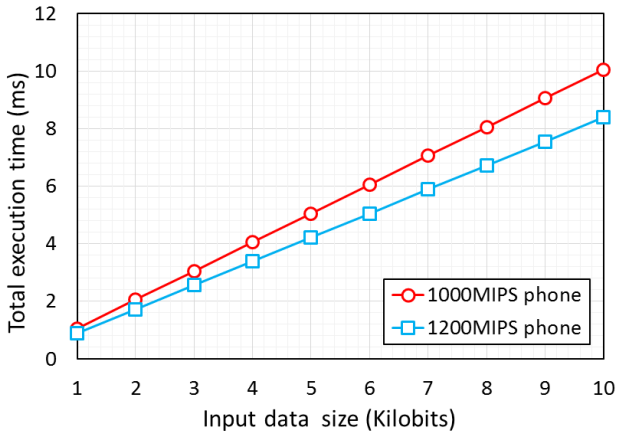

Figure 5-38. The effects of increasing input data size on the execution time of the EMCC scheme on various mobile devices.

The figure shows that as the input data size of a task increases, the execution time increases as more time is taken to execute more instructions within the MEC server. On average, for the same input data, the mobile devices with the computation capability of 1000MIPS and 1200MIPS have the total execution time of 5.6ms and 4.6ms respectively since the greater the computation capability, the faster the execution time.

Figure 5-39 shows the effects of increasing the uplink datarate on the execution time of the EMCC scheme on various mobile devices. The figure shows that for all the mobile devices, as the uplink data rate increases the total execution time decreases since more data is transmitted through the MRRH or PRRH over a short period of time to the MEC server of the BS cloud.

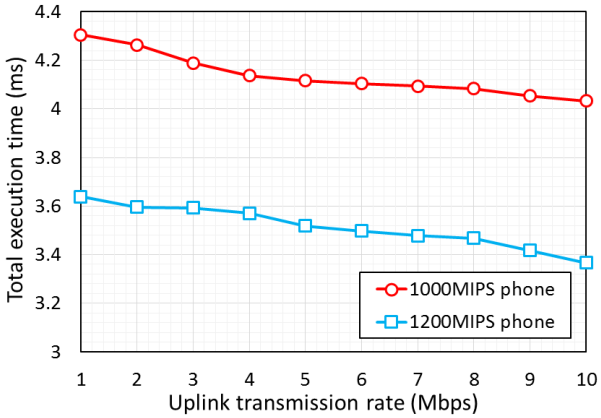

Figure 5-39. The effects of increasing uplink datarate on the execution time of the EMCC scheme on various mobile devices.

5.6.5.2 Effects on energy consumption in the mobile device

Figure 5-40 shows the average energy consumption on the mobile device for all the schemes. On average, the local computation, EECO and EMCC consumes 8.55J, 5.6J and 3.64J respectively. The EMCC scheme perfoms better than the local computation and the EECO schemes by saving 57% and 35% of energy respectively.

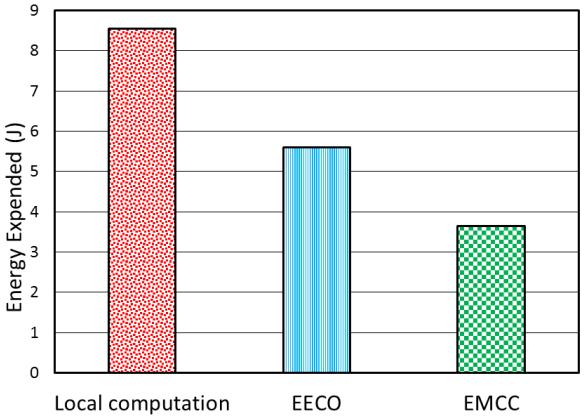

Figure 5-40. Average energy consumption on the mobile device for all the schemes.

The proposed EMCC perfoms better than the local computation scheme by 57% because in local computation, more energy is consumed since the mobile device has lower computation capability than the MEC server as such it will

take more time to execute the task which will consume a lot of energy. The proposed EMCC perfoms better than the EECO scheme by 35% because in EECO, tasks are executed in the MEC server but not in parallel whereas in EMCC, tasks are executed in parallel by the VMs in the MEC servers and this will reduce the time of execution which will results in lower energy consumption. Also, the CRE user association in PRRHs in EMCC increases the transmission datarate for the mobile device which reduces the time of transmission and in turn reduce the transmission power energy consumption.

Figure 5-41 shows the effects of increasing the number of mobile devices within RRH coverage area on the total power consumption in a single mobile device. On average, the energy consumption for Local computation, EECO and EMCC is 8.6W, 5.6W and 5W respectively. The energy consumption of

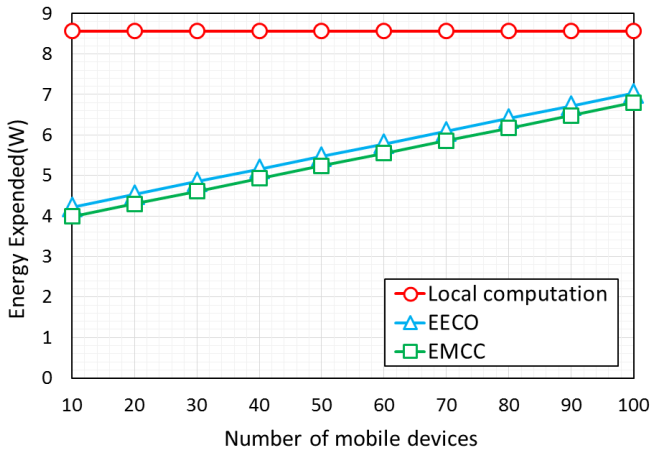

Figure 5-41. The effects of increasing the number of mobile devices on the total power consumption in a single mobile device.

the Local computation scheme is constant because local computation of tasks is not affected by the number of users in the network, to be precise, local computation is not affected by the uplink data rate and interference. But for the EECO and the EMCC, as the number of user's increases, the energy

consumed in the mobile device increases since more devices share the bandwidth which causes the uplink data rate of the mobile device to be lower and also increased interference with more users limit the uplink datarate.

5.7 Concluding Remarks

This chapter describes concluding remarks on the proposed scenarios. In Scenario 1, the results of the proposed H-C-RAN switch off scheme were presented where a baseline CAC mechanism is implemented which were compared with other schemes (LTE-A HetNet scheme, RANaaS and F-PDSO) and the results have shown that the proposed scheme switches off more number of PRRHs than all the other schemes at acceptable QoS levels while achieving significant SINR. In scenario 2, the results of the proposed workload consolidation schemes SA and GA were presented and results have shown that both the proposed schemes compared to other schemes (RANaaS, FBP, FFD, FF and NF) perform better in terms of number of GPPs switched off, statistical multiplexing gain, and GPP utilisation. The proposed SA performs better than GA in all aspects as such it is selected as the best scheme suitable for the proposed framework. To further improve performance, the H-C-RAN PRRH switch off scheme in scenario 1 was combined with the two proposed workload consolidation schemes of SA and GA which are called H-C-RAN SA and H-C-RAN GA in scenario 3 and the results show that the combination has improved performance in terms of energy efficiency, power consumption and network throughput while maintaining the required QoS. Also, the H-C-RAN SA scheme out performs the H-C-RAN GA scheme. Advanced CAC called Fuzzy CAC with pre-emption was proposed in scenario 4 within the H-C-RAN

SA scheme which was chosen for its best performance. Three schemes were compared for performance evaluation; i) baseline CAC scheme which have been used in scenarios 1 to 3, ii) proposed Fuzzy CAC without pre-emption iii) proposed Fuzzy CAC with pre-emption. The proposed Fuzzy CAC with pre-emption scheme outperformed other schemes in terms of blocking probability and GPP utilisation. Finally in scenario 5, the results for saving energy in the mobile device were presented. The proposed EMCC scheme perfoms better than the Local computation and the EECO scheme with minimum execution time and reduced energy consumption within the mobile device.

6 Conclusion and Future Work

6.1 Conclusion

The 5G RAN system will consist of billions of devices connected to the RAN system running a plethora of applications. Also the 5G RAN system will consists of HetNets system where small cell BSs are densely deployed on top of the macro cells. Based on this features, the 5G systems will suffer the following deficits;

- High-energy consumption in the RAN due to deployment of dense small cells in HetNets.
- High-energy consumption in mobile devices due to running computer intensive applications like gamming and augmented reality.
- High cost of BS equipment.
- Traffic congestion due to billions of devices connected to the 5G network.
- Underutilisation of BS equipment especially during low traffic periods where BS equipment are not shared.
- Negative environmental impact of dense BS equipment with high CO_2 emissions.

The focal objectives of this research study is to design an energy efficient framework for 5G HetNet C-RAN based system which aim at minimising energy consumption and at the same time maximising resource utilisation while maintaining QoS. The proposed framework consist of various functionalities which are summarised in the novelties below.

6.1.1 Summarised key Contributions and Simulation Results Summary

6.1.1.1 Novel computational-resource-aware power consumption model

- In order to accurately model the power usage of the framework, a new power consumption model for 5G C-RAN based on the component based model from the EARTH model has been proposed.
- The proposed model breaks down a BS into separate functions which are RF function, fronthaul and BS cloud power consumption. Each of these functions have their individual power models. Previous BS power consumption models cannot be used since in C-RAN, baseband processing power, cooling and housing are shared in the BS cloud and the RF and the BBU no longer form one BS but are actually separated.
- The separated RF function is called the RRH and its proposed power model consists of the number of antennas, the RRH traffic load, the maximum transmission power and the RF power. For the fronthaul, the power model was also presented. In the BS cloud, a standard GPP power model was adopted which scales linearly with the GPP CPU utilisation.
- A novel model for converting user traffic from the coverage area to baseband CPU load to be processed in the BS cloud was derived which was dependent on the number on antennas in mobile device, the coding rate, the modulation scheme used and the number of PRB's used.

6.1.1.2 Novel dynamic centralised small BS switch off scheme

- Designing a dynamic centralised small PRRH switch off mechanism at the radio side of Hetnet 5G C-RAN while insuring minimal dropping probability.
- Only the small cells PRRHs are switched off while the macro cells MRRH are kept active all the time to maintain coverage and minimise outage.
- The PRRHs are switched off based on the proposed utility function that is dependent on various factors such as total rate of served users, PRRH's power consumption, PRRH traffic load and the interference from other RRHs. The PRRHs with the lowest utility values are switched off first with their users successfully handed over to either the nearby RRHs or MRRH.
- The results show that the proposed PRRH switch off scheme meets the QoS requirements (2.7% blocking probability) by not exceeding the threshold blocking probability of 5%. Also the proposed scheme switches off at most 97% of the PRRHs compared to other schemes. In addition, the proposed scheme has improved SINR of up to 8dB.

6.1.1.3 Novel cloud computing based baseband workload consolidation

- The thesis proposed a new energy-efficient scheme at the BS cloud of 5G CRAN through cloud computing based baseband workload consolidation technique for minimising energy consumption.
- The traditional BBUs are replaced with virtualised GPPs due to their lower cost and programmability features for processing any baseband

signal. The rationale is to reduce the number of GPPs used by migrating vBBUs from less loaded physical GPPs to other GPPs and switching off idle GPPs during low traffic periods.

- Novel SA and GA algorithms are finally proposed for reducing the number of active GPPs and hence energy consumption in the BS cloud. The GA and the SA switche off more number of GPPs than all the other schemes with high utilisation of GPPs at marginal execution time.
- The results show that the combination of the H-C-RAN PRRH switch off scheme with the baseband consolidation schemes (GA and SA) termed H-C-RAN SA and H-C-RAN GA further enhances performance with improved energy efficiency by a factor of 10 and 9 respectively, reduced power consumption by up to 53% and 45% respectively, and improved network throughput by 53% and 45% respectively while maintaining the required QoS. The H-C-RAN SA performs better than the H-RAN GA with 8% energy reductions as such H-C-RAN SA is chosen as a scheme suitable for the workload consolidation.

6.1.1.4 Novel CAC framework for 5G C-RAN

- A novel CAC scheme for preventing traffic congestion and improving utilisation in the BS cloud is also proposed.
- The proposed CAC scheme incorporates fuzzy logic controller as the decision maker and pre-emption technique where during congestion, delay tolerant NRT low priority connections are pre-empted and outsourced to a public C-RAN cloud with a pricing penalty to

accommodate the RT connections in the private operator C-RAN, a technique called cloud bursting.

- The results show that the proposed CAC scheme satisfies the QoS with blocking probability of 0.7% which is within the acceptable limits. The proposed CAC scheme has improved GPP utilisation by 88%, improved the operator revenue by 3.5% and improved network throughput by 30% but with a penalty of increased response time of an extra 0.4ms.

6.1.1.5 Novel MEC framework for saving energy in mobile devices

- Proposed a novel MEC framework for saving energy in mobile devices within 5G C-RAN, where cloud computing capabilities are provided in close proximity to mobile devices to offload some processing requirements.
- In the proposed framework, an application from the mobile device is partitioned into modules/tasks which are either executed on the local device or offloaded and executed in parallel at the BS cloud while considering the transmission power, QoS, SoC of the mobile battery and mobile device CPU load.
- The CRE user association employed enables users within the coverage to be mostly associated with the PRRH than the MRRH as such improving the transmission data rate when executing the tasks in the BS cloud which reduces delays.
- The results show that the proposed EMCC scheme has low execution time which have decreased by 63% compared to the Local computation scheme. Moreover, the proposed EMCC scheme has significantly

reduced energy consumption within the mobile device by 57% and 35% compared to the local computation and the EECO schemes respectively.

6.2 Future Work

The work presented in this research work provides the foundation which may be further extended and studied. Some of these potential research directions are discussed in the following subsections.

6.2.1 Improving the BS Switch off Scheme

In the proposed BS switch off scheme, only the PRRHs are switched off while the MRRHs maintains coverage. This can be improved by also switching off the MRRHs which will further improve energy efficiency in the network but the trade-off between energy efficiency and QoS in terms of blocking probability and signal outage should be taken into consideration.

6.2.2 Improving BS Cloud Baseband Workload Consolidation Scheme.

In this thesis, one dimension bin packing was presented where only the CPU factor was considered. This can be improved by incorporating multidimensional bin packing by considering various resources such as RAM, storage, bandwidth, etc. Also, in this thesis, the GPPs are assumed to be of the same capacity but GPPs can be heterogeneous, with various resources. The BS cloud cooling power in the proposed scheme is considered to be fixed but this can be improved for the cooling power to linearly scale with the number of GPPs such that during low traffic, the cooling power is reduced.

6.2.3 Heterogeneous RAN Technologies

In this thesis, only the 5G RAN technology was considered based on LTE-A. The 5G RAN system will co–exist with other RAN technologies such as WiFi, 2G, 3G, 4G and WIMAX. The proposed framework could be improved to incorporate all this technologies which will be possible due the radio defined nature of the GPPs which can host any RAN technology. This will enable smooth handover to lower technologies where the 5G signal is poor. Incorporating 5G with WiFi will improved the transmission rate which will improved the energy efficiency of the proposed MEC framework which is limited by the uplink transmission data rate.

6.2.4 Energy Efficiency in the Fronthaul

In the proposed framework, only fibre was considered as the fronthaul connectivity and saving energy in the fronthaul have not been considered. This can be improved by devising novel schemes for saving energy in the fronthaul by using a mixture of fronthaul technologies and proposing some novel fronthaul data compression technics that can improve bandwidth efficiency and reduce delay.

REFERENCES

[1] Huawei, "5G: a technology vision," *White Paper,* 2013.

[2] I. Chih-Lin, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward green and soft: a 5G perspective," *IEEE Communications Magazine,* vol. 52, pp. 66-73, 2014.

[3] I. Hwang, B. Song, and S. S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Communications Magazine,* vol. 51, pp. 20-27, 2013.

[4] L. Tian, C. Liu, Y. Wan, Y. Zhou, and J. Shi, "Energy efficiency analysis of base stations in centralized radio access networks," in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2015, pp. 133-136.

[5] S. Armour, T. O'Farrell, S. Fletcher, A. Jeffries, D. Lister, S. McLaughlin*, et al.*, "Green radio: Sustainable wireless networks," *Mobile VCE,* 2009.

[6] C. R. Murthy and C. Kavitha, "A survey of green base stations in cellular networks," *International Journal of Computer Networks and Wireless Communications (IJCNWC),* vol. 2, pp. 232-236, 2012.

[7] X. Shen, "Green wireless communication networks [Editor's note]," *IEEE Network,* vol. 27, pp. 2-3, 2013.

[8] G. Fettweis and E. Zimmermann, "ICT energy consumption-trends and challenges," in *Proceedings of the 11th International Symposium on Wireless Personal Multimedia Communications*, 2008, p. 6.

[9] Cisco, "Cisco Visual Network Index: Global Mobile Data Traffic Forecast Update 2015-2016.," in *White paper*, 2013.

[10] Stuttgart. (2016, July). *Digital Agenda: EU research breakthrough will cut 4G / LTE mobile network energy use in half*. Available: http://europa.eu/rapid/press-release_MEMO-12-327_en.htm?locale=en [Accessed: July 2017]

[11] C. Mobile, "C-RAN: the road towards green RAN," *White Paper,* vol. 2, 2011.

[12] A. S. Alam, "Scalable base station switching framework for green cellular networks," The Open University, 2015.

[13] Nokia. (2016, July). *Flatten network energy consumption*. Available: https://networks.nokia.com/innovation/technology-vision/flatten-total-energy-consumption [Accessed: July 2016]

[14] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," in *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, 2011, pp. 305-316.

[15] C. Peng, S.-B. Lee, S. Lu, H. Luo, and H. Li, "Traffic-driven power saving in operational 3G cellular networks," in *Proceedings of the 17th annual international conference on Mobile computing and networking*, 2011, pp. 121-132.

[16] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *IEEE transactions on wireless communications,* vol. 12, pp. 2126-2136, 2013.

[17] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary user behavior in cellular networks and implications for dynamic spectrum access," *IEEE Communications Magazine,* vol. 47, pp. 88-95, 2009.

[18] A. Duda and C. J. Sreenan, "Challenges for quality of service in next generation mobile networks," *Parameters,* vol. 4, p. 3, 2003.

[19] T. S. Rappaport, *Wireless communications: principles and practice* vol. 2: Prentice Hall PTR New Jersey, 1996.

[20] K. Osang, "Mobile Broadband Market Trends and Insight," presented at the ITU Workshop on Bridging the Standardisation Gap, Vientiane, 2012.

[21] P. Sharma, "Evolution of mobile wireless communication networks-1G to 5G as well as future prospective of next generation communication network," *International Journal of Computer Science and Mobile Computing,* vol. 2, pp. 47-53, 2013.

[22] J. Singh, "Comprehensive Study of Mobile Networks," *International Journal of Innovative Research in Computer and Communication Engineering,* vol. 1, 2013.

[23] C. Patil, R. Karhe, and M. Aher, "Review on Generations in Mobile Cellular Technology," *International Journal of Emerging Technology and Advanced Engineering,* vol. 2, 2012.

[24] A. Gohil, H. Modi, and S. K. Patel, "5G technology of mobile communication: A survey," in *International Conference on Intelligent Systems and Signal Processing (ISSP), 2013*, pp. 288-292.

[25] T. Nakamura, A. Benjebbour, Y. Kishiyama, S. Suyama, and T. Imai, "5G radio access: Requirements, concept and experimental trials," *IEICE Transactions on Communications,* vol. 98, pp. 1397-1406, 2015.

[26] Wikipedia. (2016, January). *5G*. Available: https://en.wikipedia.org/wiki/5G [Accessed: January 2016]

[27] Huawei-Technologies. (2017, July). *5G Network Architecture: A high level pespective*. Available:http://www.huawei.com/minisite/hwmbbf16/insights/5 G-Nework-Architecture-White paper-en.pdf [Accessed: July 2016]

[28] 4IPNet. (2016, May). *All in one WIFI solution*. Available: http://4ipnet.blogspot.co.uk/2011/09/4ipnets-3g-offload-solution.html [Accessed: July 2017]

[29] Qualcomm. (2014, February). *How to minimize your app's power consumption*. Available: https://www.slideshare.net/QualcommDeveloper Network/69-minimize-powerconsumptioninappsschwartz918gg67 [Accessed: February 2017]

[30] M. Hubbard, "Taming the smartphone power consumption vicious cycle." *Microwave Journal*,vol. 55, pp. 92-96, 2012.

[31] S. Nanba and A. Agata, "A new IQ data compression scheme for front-haul link in centralized RAN," in *2013 IEEE 24th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC Workshops),* 2013, pp. 210-214.

[32] P. Rost and A. Prasad, "Opportunistic Hybrid ARQ—Enabler of Centralized-RAN Over Nonideal Backhaul," *Wireless Communications Letters, IEEE,* vol. 3, pp. 481-484, 2014.

[33] S. Namba, T. Warabino, and S. Kaneko, "BBU-RRH switching schemes for centralized RAN," in *2012 7th International ICST Conference on Communications and Networking in China (CHINACOM)* , 2012, pp. 762-766.

[34] M. Qian, Y. Wang, Y. Zhou, L. Tian, and J. Shi, "A super base station based centralized network architecture for 5G mobile communication systems," *Digital Communications and Networks,* vol. 1, pp. 152-159, 2015.

[35] Airvana. (2016, June). *Onecell architecture*. Available: http://www.airvana.com/products/enterprise/onecell-architecture/ [Accessed: February 2017]

[36] G. Auer, O. Blume, V. Giannini, I. Godor, M. Imran, Y. Jading*, et al.*, "D2. 3: Energy efficiency analysis of the reference systems, areas of improvements and target breakdown," *EARTH,* 2010.

[37] O. Arnold, F. Richter, G. Fettweis, and O. Blume, "Power consumption modeling of different base station types in heterogeneous cellular networks," in *Future Network and Mobile Summit, 2010*, pp. 1-8.

[38] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, H. Holtkamp*, et al.*, "Flexible power modeling of LTE base stations," in *2012 IEEE Wireless Communications and Networking Conference (WCNC)*, 2012, pp. 2858-2862.

[39] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger*, et al.*, "Cloud RAN for Mobile Networks; A Technology Overview," *IEEE Communications Surveys & Tutorials,* vol. 17, pp. 405-426, 2015.

[40] G. Kardaras and C. Lanzani, "Advanced multimode radio for wireless & mobile broadband communication," in *Wireless Technology Conference*, 2009, pp. 132-135.

[41] A. Maeder, M. Lalam, A. De Domenico, E. Pateromichelakis, D. Wübben, J. Bartelt*, et al.*, "Towards a flexible functional split for cloud-RAN networks," in *2014 European Conference on Networks and Communications (EuCNC)*, 2014, pp. 1-5.

[42] Alcatel-Lucent, "Leveraging VDSL2 for mobile backhaul," in *white paper*, 2010.

[43] Next-Generation-Mobile-Networks-Alliance(NGMN), "Small Cell Backhaul Requirements," June 2012.

[44] H. J. and E. J., "Microwave capacity evolution," in *Ericsson Review*, 2011, pp. 22-27.

[45] Y. Li, "E-band radios for LTE/LTE-Advanced mobile backhaul," in *2010 Workshop on Integrated Nonlinear Microwave and Millimeter-Wave Circuits (INMMIC)*, 2010, pp. 84-84.

[46] T. Sigwele, P. Pillai, and Y. F. Hu, "Call admission control in cloud radio access networks," in *2014 International Conference on Future Internet of Things and Cloud (FiCloud)*, 2014, pp. 31-36.

[47] L. Chen, H. Jin, H. Li, J.-B. Seo, Q. Guo, and V. Leung, "An energy efficient implementation of C-RAN in HetNet," in *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, 2014, pp. 1-5.

[48] A. Checko. (2016, Jan). *Cloud Ran Fronthaul*. Available: http://www.ict-ijoin.eu/wp-content/uploads/2015/03/3b_Checko_C-RAN-FH.pdf [Accessed: Feb 2017]

[49] A. Li, Y. Sun, X. Xu, and C. Yuan, "Joint remote radio head selection and user association in cloud radio access networks," in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2016, pp. 1-6.

[50] QTS. (2016, August). *Virtualisation: What It Is, What Types There Are & How It Benefits Companies*. Available: http://www.qtsdatacenters.com/resources/blog/2013/02/28/virtualisation-what-it-is-what-types-there-are-how-it-benefits-companies [Accessed: August 2016]

[51] P. T. Joy and K. P. Jacob, "Cooperative Caching Framework for Mobile Cloud Computing," *arXiv preprint,* 2013.

[52] K. Zhang, Y. Mao, S. Leng, Q. Zhao, L. Li, X. Peng*, et al.*, "Energy-Efficient Offloading for Mobile Edge Computing in 5G Heterogeneous Networks," *IEEE Access,* vol. 4, pp. 5896-5907, 2016.

[53] J. K. Zao, T. T. Gan, C. K. You, S. J. R. Méndez, C. E. Chung, Y. Te Wang*, et al.*, "Augmented brain computer interaction based on fog computing and linked data," in *2014 International Conference on Intelligent Environments (IE)*, 2014, pp. 374-377.

[54] A. Ahmed and E. Ahmed, "A survey on mobile edge computing," in *10th International Conference on Intelligent Systems and Control (ISCO)*, 2016, pp. 1-8.

[55] R. R. Sarukkai and A. Mendhekar, "Method and apparatus for accessing targeted, personalized voice/audio web content through wireless devices," ed: Google Patents, 2004.

[56] H. Liu, F. Eldarrat, H. Alqahtani, A. Reznik, X. de Foy, and Y. Zhang, "Mobile Edge Cloud System: Architectures, Challenges, and Approaches," *IEEE Systems Journal,* 2017.

[57] C. Han, T. Harrold, S. Armour, I. Krikidis, S. Videv, P. M. Grant*, et al.*, "Green radio: radio techniques to enable energy-efficient wireless networks," *IEEE communications magazine,* vol. 49, 2011.

[58] Z. Dietrich. (2017, January). *EARTH: energy aware radio and networks technologies*. Available: https://www.ict-earth.eu/ [Accessed: Jan 2017]

[59] Alcatel-Lucent. (2016, December). *Green Touch*. Available: http://www.greentouch.org/ [Accessed: Dec 2016]

[60] A. Luis and V. Christos. (2016, November). *greenet - An initial training on green wireless networks* . Available: http://www.fp7-greenet.eu/default.php [Accessed: Nov 2016]

[61] NGMN, "Suggestions on potential solutions to C-RAN by NGMN alliance," in *P-CRAN, centralized processing, collaborative radio, real-time cloud computing clean ran system*, 2013.

[62] European-Commission. (2017, Jan). *Mobile Cloud Networking (MCN) Project*. Available: http://www.mobile-cloud-networking.eu/site [Accessed: Jan 2017]

[63] European-Commission. (2017, Jan). *FP7 Project High Capacity Network Architecture With Remote Radio Heads and Parasitic Antenna Arrays (HARP)*. Available: http://www.fp7-harp.eu/ [Accessed: Jan 2017]

[64] European-Commission. (2017, Jan). *Interworking and JOINt design of an open access and backhaul network architecture for small cells based on cloud networks*. Available: http://www.ict-ijoin.eu/ [Accessed: Jan 2017]

[65] European-Commission. (2017, Feb). *Connectivity Management for Energy Optimized Wireless Dense networks (CROWD)*. Available: http://www.ict-crowd.eu/ [Accessed: Feb 2017]

[66] Z. Ghebretensaé, K. Laraqui, S. Dahlfort, F. Ponzini, L. Giorgi, S. Stracca*, et al.*, "Transmission solutions and architectures for heterogeneous networks built as C-RANs," in *7th International ICST Conference on Communications and Networking in China (CHINACOM)*, 2012, pp. 748-752.

[67] C. Liu, K. Sundaresan, M. Jiang, S. Rangarajan, and G.-K. Chang, "The case for re-configurable backhaul in cloud-RAN based small cell networks," in *INFOCOM, 2013 Proceedings IEEE*, 2013, pp. 1124-1132.

[68] Korea-Telecom. (2017, March). *Rethink Wireless, Korea Telecom Plans World's First Commercial Cloud-RAN [Online]*. Available: http://rethink-wireless.com/2011/12/08/korea-telecom-plans-worlds-commercial-cloud-ran [Accessed: March 2017]

[69] ZTECoporation, "ZTE green technology innovations," in *ZTE white paper*, 2011 .

[70] S.-E. Elayoubi, L. Saker, and T. Chahed, "Optimal control for base station sleep mode in energy efficient radio access networks," in *INFOCOM, 2011 Proceedings IEEE*, 2011, pp. 106-110.

[71] L. Saker, S.-E. Elayoubi, and T. Chahed, "Minimizing energy consumption via sleep mode in green base station," in *2010 IEEE Wireless Communication and Networking Conference*, 2010, pp. 1-6.

[72] P. Dini, M. Miozzo, N. Bui, and N. Baldo, "A model to analyze the energy savings of base station sleep mode in LTE HetNets," in *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing*, 2013, pp. 1375-1380.

[73] J. Wu, Y. Wu, S. Zhou, and Z. Niu, "Traffic-aware power adaptation and base station sleep control for energy-delay tradeoffs in green cellular networks," in *Global Communications Conference (GLOBECOM), 2012 IEEE*, 2012, pp. 3171-3176.

[74] S. Zhou, J. Gong, Z. Yang, Z. Niu, and P. Yang, "Green mobile access network with dynamic base station energy saving," in *ACM MobiCom*, 2009, pp. 10-12.

[75] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "On the effectiveness of single and multiple base station sleep modes in cellular networks," *Computer Networks,* vol. 57, pp. 3276-3290, 2013.

[76] J. Wu, S. Jin, L. Jiang, and G. Wang, "Dynamic switching off algorithms for pico base stations in heterogeneous cellular networks," *EURASIP Journal on Wireless Communications and Networking,* vol. 2015, pp. 1-18, 2015.

[77] T. Zhao, J. Wu, S. Zhou, and Z. Niu, "Energy-delay tradeoffs of virtual base stations with a computational-resource-aware energy consumption model," in *IEEE International Conference on Communication Systems (ICCS)*, 2014, pp. 26-30.

[78] N. Saxena, A. Roy, and H. Kim, "Traffic-aware cloud RAN: a key for green 5G networks," *IEEE Journal on Selected Areas in Communications,* vol. 34, pp. 1010-1021, 2016.

[79] S. Namba, T. Matsunaka, T. Warabino, S. Kaneko, and Y. Kishi, "Colony-RAN architecture for future cellular network," in *Future Network & Mobile Summit (FutureNetw), 2012*, 2012, pp. 1-8.

[80] M. Qian, W. Hardjawana, J. Shi, and B. Vucetic, "Baseband processing units virtualisation for cloud radio access networks," *IEEE Wireless Communications Letters,* vol. 4, pp. 189-192, 2015.

[81] A. Al-Dulaimi, S. Al-Rubaye, and Q. Ni, "Energy efficiency using cloud management of LTE networks employing fronthaul and virtualized baseband processing pool," 2016.

[82] D. Sabella, A. De Domenico, E. Katranaras, M. A. Imran, M. Di Girolamo, U. Salim*, et al.*, "Energy Efficiency benefits of RAN-as-a-

Service concept for a cloud-based 5G mobile network infrastructure," *IEEE Access,* vol. 2, pp. 1586-1597, 2014.

[83] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata*, et al.*, "Coordinated multipoint transmission and reception in LTE-advanced: deployment scenarios and operational challenges," *IEEE Communications Magazine,* vol. 50, pp. 148-155, 2012.

[84] P.-R. Li, T.-S. Chang, and K.-T. Feng, "Energy-efficient power allocation for distributed large-scale MIMO cloud radio access networks," in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, 2014, pp. 1856-1861.

[85] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Transactions on Networking,* vol. 24, pp. 2795-2808, 2016.

[86] V. N. Ha, L. B. Le, and N.-D. Dào, "Energy-efficient coordinated transmission for cloud-RANs: Algorithm design and trade-off," in *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, 2014, pp. 1-6.

[87] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Communications,* vol. 21, pp. 126-135, 2014.

[88] M. Khan, R. Alhumaima, and H. Al-Raweshidy, "Reducing energy consumption by dynamic resource allocation in C-RAN," in *European Conference on Networks and Communications (EuCNC)*, 2015, pp. 169-174.

[89] Z. Kong, J. Gong, C.-Z. Xu, K. Wang, and J. Rao, "ebase: A baseband unit cluster testbed to improve energy-efficiency for cloud radio access network," in *2013 IEEE International Conference on Communications (ICC)*, 2013, pp. 4222-4227.

[90] M. Y. Lyazidi, N. Aitsaadi, and R. Langar, "Dynamic resource allocation for Cloud-RAN in LTE with real-time BBU/RRH assignment," in *Communications (ICC), 2016 IEEE International Conference on*, 2016, pp. 1-6.

[91] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches," *Wireless communications and mobile computing,* vol. 13, pp. 1587-1611, 2013.

[92] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE*

Journal on Selected Areas in Communications, vol. 34, pp. 3590-3605, 2016.

[93] L. Chen, S. Zhou, and J. Xu, "Energy Efficient Mobile Edge Computing in Dense Cellular Networks," *arXiv preprint,* 2017.

[94] M. Deng, H. Tian, and X. Lyu, "Adaptive sequential offloading game for multi-cell Mobile Edge Computing," in *23rd International Conference on Telecommunications (ICT)*, 2016, pp. 1-5.

[95] M. T. Beck, S. Feld, A. Fichtner, C. Linnhoff-Popien, and T. Schimper, "ME-VoLTE: Network functions for energy-efficient video transcoding at the mobile edge," in *18th International Conference on Intelligence in Next Generation Networks (ICIN)*, 2015, pp. 38-44.

[96] H. Zhang, C. Jiang, J. Cheng, and V. C. Leung, "Cooperative interference mitigation and handover management for heterogeneous cloud small cell networks," *IEEE Wireless Communications,* vol. 22, pp. 92-99, 2015.

[97] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam*, et al.*, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Communications,* vol. 18, pp. 10-21, 2011.

[98] P. Ökvist and A. Simonsson, "LTE HetNet trial-range expansion including micro/pico indoor coverage survey," in *Vehicular Technology Conference (VTC Fall), 2012 IEEE*, 2012, pp. 1-5.

[99] D. López-Pérez, A. Ladanyi, A. Jüttner, H. Rivano, and J. Zhang, "Optimisation method for the joint allocation of modulation schemes, coding rates, resource blocks and power in self-organizing LTE networks," in *INFOCOM, 2011 Proceedings IEEE*, 2011, pp. 111-115.

[100] M. Imran and E. Katranaras, "Energy efficiency analysis of the reference systems, areas of improvements and target breakdown. ICT-EARTH Project, Deliverable D2. 3, EC-IST Office, Brussels, Belgium (January 2011)," ed.

[101] S. Tombaz, P. Monti, K. Wang, A. Vastberg, M. Forzati, and J. Zander, "Impact of backhauling power consumption on the deployment of heterogeneous mobile networks," in *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, 2011, pp. 1-5.

[102] P. Monti, S. Tombaz, L. Wosinska, and J. Zander, "Mobile backhaul in heterogeneous network deployments: Technology options and power consumption," in *14th International Conference on Transparent Optical Networks (ICTON),* , 2012, pp. 1-7.

[103] T. Werthmann, H. Grob-Lipski, and M. Proebster, "Multiplexing gains achieved in pools of baseband computation units in 4G cellular networks," in *PIMRC*, 2013, pp. 3328-3333.

[104] S. Kirkpatrick, "Optimisation by simulated annealing: Quantitative studies," *Journal of statistical physics,* vol. 34, pp. 975-986, 1984.

[105] R. Rao and S. Iyengar, "Bin-packing by simulated annealing," *Computers & Mathematics with Applications,* vol. 27, pp. 71-82, 1994.

[106] E. Falkenauer and A. Delchambre, "A genetic algorithm for bin packing and line balancing," in *IEEE International Conference on Robotics and Automation*, 1992, pp. 1186-1192.

[107] W. H. Tranter, T. S. Rappaport, K. L. Kosbar, and K. S. Shanmugan, *Principles of communication systems simulation with wireless applications* vol. 1: Prentice Hall New Jersey, 2004.

[108] W. T. Kasch, J. R. Ward, and J. Andrusenko, "Wireless network modeling and simulation tools for designers and developers," *IEEE Communications Magazine,* vol. 47, pp. 120-127, 2009.

[109] J. R. Norris, *Markov chains*: Cambridge university press, 1998.

[110] FIT4Green. (2015, Sept). *Presentation of full-featured federated energy consumption models*. Available: http://www.fit4green.eu/sites/default/files/attachments/documents/D3.3_final.pdf. [Accessed: Sept 2016]

[111] Intel. (2015, Aug). *Intel Xeon Processor 5500 series [Online]*. Available: http://download.intel.com/support/processors/xeon/sb/xeon_5500.pdf [Accessed: Aug 2016]

[112] X. Wang, S. Thota, M. Tornatore, H. S. Chung, H. H. Lee, S. Park*, et al.*, "Energy-Efficient Virtual Base Station Formation in Optical-Access-Enabled Cloud-RAN," *IEEE Journal on Selected Areas in Communications,* vol. 34, pp. 1130-1139, 2016.

[113] A. Hussain. (2017, June). *How to dimesion user traffic in 4G*. Available: https://www.slideshare.net/althafhussain1023/how-to-dimension-user-traffic-in-lte, [Accessed: June 2017].

[114] H. Gupta, A. V. Dastjerdi, S. K. Ghosh, and R. Buyya, "iFogSim: A Toolkit for Modeling and Simulation of Resource Management Techniques in Internet of Things, Edge and Fog Computing Environments," *arXiv preprint,* 2016.

[115] S. N. Shahab. (2016, June). *Any advice on the Maximum SINR for LTE?*Available:https://www.researchgate.net/post/Any_advice_on_the_Maximum_SINR_ for_LTE [Accessed: Sept 2017]